문학석사 학위논문

Investigating the Feasibility of Using Interactive Phone Communication Tasks in an EFL Speaking Assessment Context

EFL 말하기 평가 상황에서 상호작용적
전화대화 과제의 활용 가능성 연구

2013 년 8월

서울대학교 대학원

영어영문학과 어학전공

정 루 미

Investigating the Feasibility of Using Interactive Phone Communication Tasks in an EFL Speaking Assessment Context

지도 교수 이용원

이 논문을 문학석사 학위논문으로 제출함
2013 년 8월

서울대학교 대학원
영어영문학과 어학전공
정 루 미

정루미의 문학석사 학위논문을 인준함
2013년 8월

위 원 장 _____ (인)

부위원장 _____ (인)

위 원 _____ (인)

Abstract

Technology-mediated speaking assessment has become widely used in recent years due to the benefits of test delivery and score reliability. On the negative side, however, semi-direct tests have been criticized in that they lack interaction, and thus only a narrow range of language skills can be assessed. Although direct-speaking language tests will be available when an interactional component is added by utilizing mobile phones or VoIP (Voice over Internet Protocol), studies have been scant with regard to the feasibility of implementing interactional technology in a semi-direct speaking assessment. Thus, the current study empirically investigated the feasibility of using interactive mobile phone communication tasks in an EFL speaking assessment.

Forty-four Korean university students who learned English as a foreign Language were recruited to perform speaking tasks that were created by the researcher. Two types of tasks were used to examine the test-takers' performances: a role-play type of non-interactive tasks featuring voicemail (monologic); and an interactive task in the form of mobile phone communication, where dialogic-level interaction takes places with one interlocutor. The test-takers' performances were evaluated by two raters, based on five analytic criteria for each task. Additionally, questionnaires were used to examine how the test-takers and the raters perceived the two different types of tasks.

Analysis of the results showed that a high degree of inter-rater

reliability was achieved, although rater-score disagreements were found in some analytic criteria such as task achievement and discourse management in the non-interactive test and interactional communication in the interactive test. To support a claim of validity, a strong positive relationship among items and tests was also found. To analyze for noteworthy patterns demonstrated in the two tasks, the length of each test-taker's speech time on each of the two tasks and the conversational turns occurring in the interactive tasks were counted. This indicated that more information regarding fluency may be revealed in the interactive tasks in that spontaneous reaction is required in real-time interactions. The analysis of the questionnaires found that thirty-eight test-takers out of forty viewed the interactive tasks as being a more authentic, realistic, and accurate measure of assessing speaking ability than the non-interactive tasks in that real-communication ability was revealed.

Overall, the interactive phone-communication tasks are a reliable and authentic measure to assess interactive speaking skills as a way to fill the gap in the semi-direct speaking format where only monologic tasks are used to elicit test-takers' speech. Further studies are needed to investigate test-takers' performances and their perceptions in various TLU domains in phone communication contexts where a variety of social relationships and functional languages are involved.

Keywords: semi-direct assessment, interaction, interactive phone-communication tasks, role-plays, target language use

Student number: 2010-20020

Table of Contents

List of Tables

List of Figures

Chapter 1

Introduction

1.1 Background and Motivation

Speaking assessment in second language learning is considered to be very important because communicative ability is a crucial element of one's success in language acquisition (Luoma, 2004). With the increased need for oral proficiency assessment, not only have researchers in the fields of second language acquisition and teaching conducted studies to understand "what constitutes speaking ability" (Fulcher, 2003), but also language testers have made various attempts to design speaking tests and rating scales that measure second language learners' speaking ability in a reliable and authentic way.

One critical consideration in assessing second language learners' speaking ability has to do with how to obtain speech samples from test-takers. Interviews, role plays, and picture description tasks are commonly-used methods to elicit speech samples from candidates. In terms of testing formats, there are several different kinds to choose from. Oral proficiency interviews and paired assessments are two direct testing formats attracting a lot of attention recently in language testing because they assess direct face-to-face conversational ability. In addition, there are a variety of semi-direct testing formats that are widely used in speaking assessment, particularly in large-scale standardized speaking assessment, due to their versatility and cost efficiency (Qian, 2009).

Semi-direct testing is rather a broad term used to refer to a wide variety

of oral proficiency assessments. In this mode of assessment, the test-taker receives input from various multimedia sources (e.g., cassette tape, compact disc, computer, and internet) and is asked to perform speaking tasks based on the input received (Qian, 2009). The test-taker's speech performance is usually recorded on a tape, on a disc, or as a digital computer file and distributed to one or more raters at different locations for scoring purposes.

The term SOPI (Simulated Oral Proficiency Interview) was coined by Stansfield and Kenyon (1988) as an alternative term for semi-direct speaking assessments in general. When the semi-direct format of oral assessment was first created in the US in the 1980s, it was mostly tape-mediated. But since then, with advances in technology, computer-based speaking tests have been introduced to semi-direct speaking assessment. Such assessments are now called Computerized Oral Proficiency Instruments (COPI), and are expected to be a more common format of speaking assessment in the future (Galaczi, 2010).

Accordingly, research into the comparability of scores obtained from computerized testing and those from conventional ones have been conducted. It was found that semi-direct test formats still measure the same skills as direct test formats do (Stansfield & Kenyon, 1992) do. High correlations were obtained between the scores from semi-direct and direct speaking assessment formats (Shohamy, 1983). Nevertheless, some studies have reported that the nature of the language in semi-direct tests is less oral-like, and the level of anxiety tends to increase in the semi-direct speaking test mode. These findings suggest that the measure and contexts used to elicit language are influential

factors on language features elicited, and thus to be evaluated.

Some researchers claimed that semi-direct speaking assessment is a more reliable method of assessment than direct in that it can be administered to test-takers in an identical way regardless of where the test is administered. It is also an efficient and practical test format for both candidates and raters because testing in this format takes in this format a short amount of time compared to other testing formats, and speech production can be stored and rated at any time. For this reason, computer-based assessment (CBA henceforth) has become a more common way to assess speaking in the last five years. In fact, CBA has gained status as an integral part of quality high-stakes testing, as seen in the examples of ETS (Education Testing Service)'s TOEFL iBT speaking test and TOEIC speaking test, Cambridge ESOL's BULATS online speaking test, Pearson's VERSANT speaking test, and the ACTFL OPIc (Oral Proficiency Interview-computer). These tests are regularly administered to a large population of test-takers worldwide for education and business purposes.

Computer-based speaking assessment usually comprises various tasks to elicit speech samples of test-takers. Each of these assessment tasks represents real-life communication situations, so that the candidates' future performance can be inferred based on their performance on the assessment task. However, computer-based speaking assessment has been criticized in terms of construct validity because only a narrow range of skills can be captured in such constrained tasks (Chun 2006). The major component of the criticism is that oral proficiency tests in a CBA format lack interactional

components to assess candidates' ability to deal with communication. O'Sullivan, Weir and Saville (2002) argued that computer-based speaking assessment can assess the ability to describe visuals and, provide personal information in the form of constructing one's own message, but it cannot capture the ability to initiate an interaction, change the topic, or terminate the interaction. Such ability is related to interactional components of speaking and can best be captured in a face to face oral proficiency interview format where interaction takes place in an authentic way.

Since CBA is task-based assessment and reflects target language use domains, tasks in communication contexts such as writing an email for writing assessment and leaving a voice message for speaking are often provided. In the current state of CB speaking assessment, a monologic type of tasks is easily available due to its logistical advantages. Among the tasks, in a leaving-a-voice-message task in particular, candidates who are given a hypothetical situation must construct their message based on the information provided in a reading text or a listening file. Although such tasks try to represent modern life, because real life depends on a various communication means, real-time conversational ability cannot be assessed well by CB assessments

Given the clear limitations of SOPI and COPI, using interactive phone communication tasks in speaking assessment seem to have immense potential in terms of improving the validity and authenticity of existing speaking tests. The interactive phone communication tasks could make it possible to capture the interactive aspect of oral communication engaged by

test-takers and at the same time represent an authentic oral task for real life communication situations. Nowadays, people increasingly rely on phone communication, particularly using mobile phones or internet phones (Wong and Waring, 2010). In order to better represent the major target language use, speaking tasks in the context of a modern technology-mediated society, should include a phone communication task involving dialogic interaction. From this perspective, test formats without interactions involved may lack authenticity and validity. Moreover, the leaving-a-voice-message type of tasks are rarely used nowadays, especially in the Korean context, which seems to make the inclusion of such tasks increasingly irrelevant in speaking tests in terms of representing target language use domain.

In addition to the two aforementioned advantages, the interactive phone communication tasks highlighted here have three noteworthy or desirable task features from the assessment perspective: (a) demand for real time processing; (b) role playing; and (c) problem-solving.

The first notable feature of phone communication tasks has to do with the real-time processing of turn-taking, which also is required for interlocutors in face-to-face interaction. Phone communication tasks require test-takers to process language input in real time through turn taking. This also means that the automaticity of language processing can be assessed in less planned, improvised, and interactional contexts. Such technology-mediated, real-time interaction happens frequently nowadays in business and academic settings where it is difficult to communicate face to face. In a sense, adopting interactive phone communications tasks as assessment tasks may represent an

effort to combine face-to-face interview and communication technology. Galaczi (2010) also predicts that such hybridization of traditional face-to-face oral assessment and communication technology can hold great potential for future development in speaking assessment.

Second, interactive phone communication requires a test-taker to take and play a certain social role as specified in the prompt when she or he interacts with an interlocutor or examiner. In other words, the test-taker's speech performance is elicited from dialogic role-play situations and evaluated through statistical and qualitative analyses. This also helps researchers to use role-plays to delineate methods for identifying potential problems and suggesting some solutions to resolve those problems that could occur between friends in phone communication situations. The interaction with the test-taker is constrained to a friend since the participants for this study are university students; variables resulting from pragmatics may be controlled if the interactive partner is constrained.

The third important feature of the tasks is that the nature of goal-oriented communication can be easily simulated in the interactive phone communication tasks. Collaborative problem-solving is a significant part of our daily experiences. Knowing how to report a problem and then solve the problem is an essential skill in order to function effectively in any society. In speaking assessment contexts, adding problem-solving components to the dialogic communication tasks can elicit meaningful speech samples from the test-takers. As mentioned previously, such a format of testing elicits speech production in a more authentic way.

Nevertheless, one can criticize that this task type may not be easily implementable in large-scale assessment because trained examiners or interlocutors are needed. This means that test-takers can be unfairly influenced by examiner biases and personality factors as well as rater characteristics. However, it should be pointed out that such criticisms are also applicable to oral proficiency interviews that are still thriving as a viable form of speaking assessment. Despite the current debates about the value of interactive phone communication tasks, there has been a lack of research to empirically examine the reliability and validity of the scores obtained from such tasks in second language speaking assessment contexts.

With such a backdrop, the main purposes of the present study are to: (a) develop prototype versions of interactive phone communication tasks, and; (b) investigate the feasibility of using such tasks in standardized speaking assessment. In this study, the reliability and validity of the newly-developed interactive tasks are evaluated in comparison with non-interactive versions of phone communication tasks widely used in semi-direct speaking assessment.

To evaluate the test-takers' performances, existing rating scales can be used as they are without revision or these scales may need to be modified to capture the interactional aspect of performance, including turn-taking and turn-giving. A detailed description of the facets of the test methods may also need to be provided to examine the validity of certain features of the tasks in fully representing the construct of oral proficiency assessment. Various psychometric and statistical indices need to be computed, which represent qualities of the scores obtained from interactive and non-interactive phone

communication tasks. These indices include inter-rater reliability indices and correlations among analytic scores within and between tasks and between task scores and other criterion scores. To supplement quantitative analyses of speaking scores, the textual analyses of the test-takers' speech samples can be done by computing and comparing the speech length for both the interactive and non-interactive tasks and analyzing conversational turns for the interactive tasks. Test-takers' and raters' feedback about the interactive tasks were elicited to find out how they perceive the two different tasks and from raters to find out how they perceive the two different tasks and their performances.

1.2 Research Questions

1. Do interactive and non-interactive tasks achieve equally acceptable levels of inter-rater reliability?
2. How do the scores from interactive tasks correlate with those from non-interactive tasks and other criterion scores?
3. Do speech samples from interactive and non-interactive tasks show some noteworthy patterns of textual characteristics?
4. How do EFL test-takers perceive interactive tasks as compared to non-interactive tasks?
5. How do raters and an interlocutor perceive test-takers' performance on an interactive task as compared to a non-interactive task?

1.3 Organization of the Thesis

The rest of the study is organized as follows. Chapter 2 provides ta heoretical background and an overview of major formats of speaking assessment and empirical studies investigated to find more valid, reliable, and authentic measure. Chapter 3 presents the method, and Chapter 4 provides results of the study including the findings from the questionnaires collected from test-takers and raters. Chapter 5 covers discussion of the results. Finally, the conclusion of the study, limitations, and suggestions for further research are provided in Chapter 6.

Chapter II

Literature Review

This chapter deals with a review of previous literature on second language speaking assessment and studies conducted to examine not only the authenticity and content validity of speaking assessments but also the reliability and validity of the scores obtained from these speaking assessments. First, Fulcher's (2003) speaking assessment framework is briefly discussed. Second, traditional face-to-face speaking assessments are presented, along with some empirical results and findings. Next, a summary of previous research in semi-direct speaking assessment will be introduced along with an overview of computer-based speaking assessment and its characteristics. Finally, previous studies on examinees performance on monologic role-play and dialogic role-play tasks will be provided in order to support the rationale for the use of an interactive task in semi-direct speaking assessment.

2.1 Theoretical Framework of Speaking Tests

Several speaking models (e.g., Fulcher, 2003; McNamara, 1996; Milanovic and Saville, 1996; Skehan, 1998) have been proposed to understand the operational definition of speaking ability in an assessment context. The theoretical framework of speaking tests in language testing has been changed and developed more concisely with the increased understanding of the speaking construct and the interaction between the test-taker's speaking

performance and other variables, resulting from the refinement of experimental design and statistical analysis measurement. McNamara's model (1996) first depicts the interaction between the test-taker's performance and the rater's rating process. Overall, this speaking model indicates that speaking performance elicited from tasks and evaluation of the performance are the two most essential components in testing speaking ability, and various interacting factors come into play.

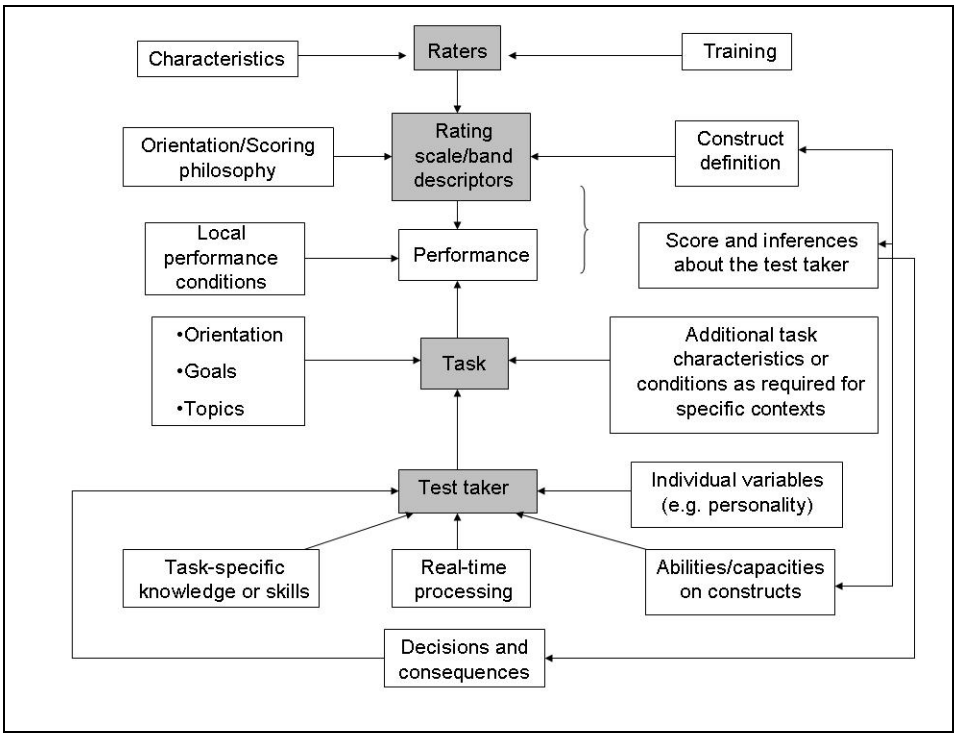


Figure 2.1 Fulcher (2003;115)'s expanded model of speaking test performance

Skehan's model (1998) further takes task characteristics and task implementation conditions into account. This model illustrates that the interactive performance, the test-taker's ability to use language in testing, and task characteristics and conditions are the main factors affecting the test score.

Fulcher's model (Figure 2.1) is an expanded version of Skehan's model in that it depicts the effects of rating scales on making inferences about the test-taker's performance. Also in his model, elements that influence tasks, including orientation, interactional relationship, goals, interlocutors, topics, and situations, are taken into consideration, along with additional task characteristics or conditions as required for specific contexts.

2.2 Direct and Semi-direct Tests

2.2.1 Key terminology

There are three types of testing modes in speaking assessment. These three modes can be categorized into "indirect," "direct," and "semi-direct," and refer to the method of test administration Indirect speaking assessment refers to the earlier versions of oral proficiency tests that were delivered mostly in the form of dictation tasks, mechanical repetition of a series of words and sentences, and pattern answers to pattern questions (Shohamy, Reves and Bejerano, 1985). Direct and semi-direct tests are the most widely used modes these days. Direct tests involve face- to- face interaction with an examiner or another candidate, while in semi-direct tests, speech samples are elicited through a series of pre-recorded prompts without interaction. Semi-direct tests are administered via tape (mostly in the past), computer or VoIP (Voice over Internet Protocol).

Scoring speech performance produced through tests is another crucial component in speaking assessment. With the scoring process taken into account, direct and semi-direct speaking test formats can be placed on a grid

with two axes: delivery and scoring, as demonstrated in Figure 1 (cited in Galaczi, 2010). Direct speaking format assessment is located in quadrant 1, where humans are involved in both delivering the test and scoring the performance. Cambridge ESOL's suite of General speaking tests and the ACTFL OPI tests are good examples of such assessment. Tests such as the TOEFL iBT speaking test and the BULATS online speaking test (the TOEIC Speaking test) are located in quadrant 2, where tests are delivered via computer, but candidates' speaking performances are scored by human raters. Quadrant 3 represents tests that are delivered and scored by computer. The Versant test, and the TOEFL iBT speaking practice tests are examples of this type. Tests that are delivered by humans but are scored by computer, represented in quadrant 4, are not currently available.

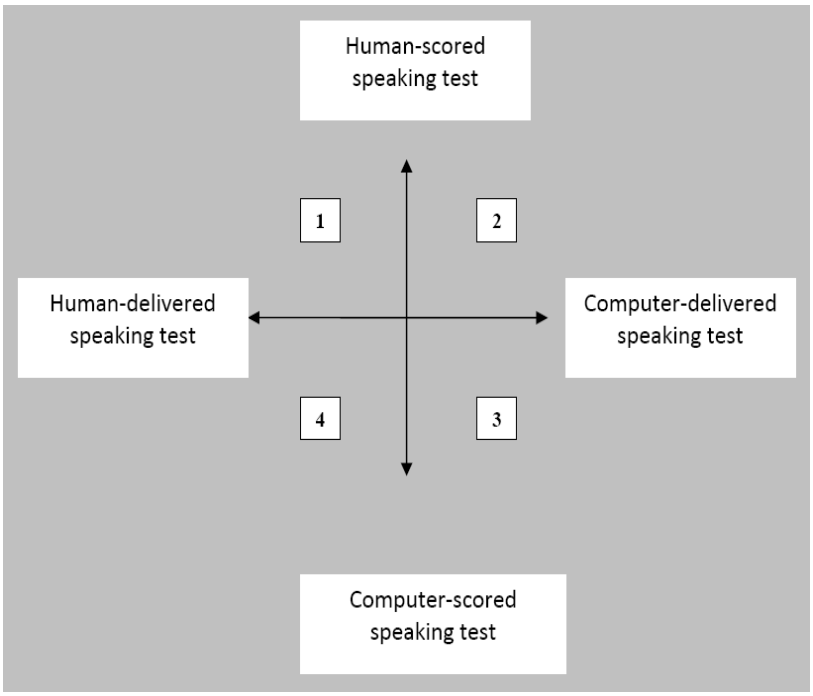


Figure 2.2 Delivery and scoring possibilities in speaking assessment (from Galaczi, 2010)

2.2.2 Traditional Face- to- face Speaking Assessment

The earlier oral proficiency assessment format was less communicative in that most testing items required performing mechanical repetitions of a series of words and sentences, and providing simple answers to given pattern questions (Shohamy, Reves and Bejerano, 1985). However, ever since communicative competence has gained attention with its emphasis on success in real-life oral communication, language testers have made an attempt to develop assessment in a more communicative and authentic way. Communicative competence refers to not only learners' knowledge about a second language but also their knowledge about how to use it appropriately (Canale and Swain, 1980; Carroll, 1983). To assess this ability, researchers suggested designing authentic assessments which can measure one's actual ability to perform the language as would be required in the real-life domain.

The oral interview format is one such assessment, in which examiners (the interviewer) and candidates (interviewee) have a conversation through interacting with each other on various topics in a context that is similar to what they might experience in their daily lives, where questions and answers are used frequently. The oral proficiency interview (henceforth, OPI) has become an important tool of speaking assessment, as it is considered to capture the true speaking ability of candidates in a communicative context. The OPI is a widely -used type of standardized interview format for assessing oral proficiency, gaining popularity for its face validity (Fulcher, 1997). In this direct assessment format, as mentioned earlier, test-takers answer

questions on a variety of topics asked by an interviewer, and their responses are evaluated based on a rating scale. The interviewer can serve as the examiner or the interlocutor, depending on the test setting. One of the major strengths of the OPI is that it can actively and appropriately capture a learner's ability to engage in a conversation. Such conversational ability is viewed as an important component of assessing L2 speakers' oral proficiency in that language ability to perform openings and closings, to establish and change topics, to hold and yield the floor, to backchannel, and to collaborate are characteristics of communicative competence that can be elicited through face to face interaction (Celce-Murcia *et al*., 1995).

However, concerns have been voiced about the construct validity of the OPI. The claim that the OPI has great potential to measure communicative competence by assessing "*speaking ability in real-life context*" (Education Testing Service, 1982: 13 cited in van Lier (1989)) was called into question. To examine the claim, van Lier (1989) first embarked on investigation into the similarities and differences between speech samples in OPIs and everyday conversation. Asymmetrical contingency, and/or pseudocontingency were observed in the OPIs while reactive and mutual contingency were found in real-life conversations. Thus, he argued that language elicited from interviews does not represent characteristics of natural conversation. That is, in interviews, examiners and candidates did not have equal rights and duties because interviewers have control over the conversation. The ability to perform openings and closings, and to establish and change topics are particularly hard to measure. Similar findings were supported in Young and

Milanovic (1992), where the interview section of the Cambridge First Certificate in English (FCE) was analyzed in terms of dominance, contingency and goal orientation. The study demonstrated that interviews are asymmetrically contingent. By accommodating and reformulating the topics, interviewers have more control over the topics discussed in the discourse. Another study conducted by Ross and Berwick (1992) reported that interlocutors had a tendency to control the conversation by nominating and reformulating the topics, and they tended to employ accommodation with less proficient examinees more frequently than they did with ones who were at an advanced proficiency level.

While a substantial body of literature regarding interaction has focused on comparing interactions in the interview context with ones found in the non-test context, or in characterizing patterns displayed in the discourse between interviewers and interviewees, some studies have been concerned with interviewers and accommodations (Ross, 1996; Brown and Lumely, 1997; Reed and Halleck, 1997). Common threads searched in this line of exploration are that interviewers have their own distinctive styles, and thus there are interviewer variations in the oral interview test, that may influence an interviewee's performance. For example, Lazaraton (1996) found that interaction which occurs in the interview format assessment was different from that which occurs in natural conversation in terms of the way interlocutors facilitate the conversation, although some supportive evidence for the OPI was also observed. In other words, some components that were present in non-assessment contexts were observed in the interviews. Eight

types of conversational techniques such as —priming topics, providing words or collaborating turns, giving evaluative feedback, repetitions and corrections of responses, slow speech along with overarticulation and pauses, stating questions that only require yes-no confirmations, drawing conclusions, and rephrasing questions —were used by the interlocutors, and provided signs of interviewer support. However, upon a closer examination, it was revealed that how interlocutors interact with candidates could be problematic since it was not consistent, and such varied and uneven interview techniques and behaviors on the part of interlocutors might influence a candidates' performance, which could lead to unfair evaluation.

Thus, traditional face-to face interviews have also been criticized in terms of score reliability. In fact, the studies concerning the impact of interviewers on ratings revealed that ratings were influenced by not only interviewers' interview skills or styles, but also by levels of functional language elicited by tasks (Shohamy, 1983; Morton etal, 1997; McNamara and Lumley, 1997; Brown and Hill, 1998). Moreover, proficiency measured by scores obtained from the interview is questionable because speaking performance can be affected by additional variables such as the setting, time, style and gender of interviewers, and the status, and role in interaction between interviewers and interviewees. Studies conducted by Shohamy (1983; 1988) found additional effects caused by variables such as testing time and interviewer. Imbalance between interviewer and interview in terms of status, and the amount of talk were reported. Due to their higher status, it is inevitable that interviewers have power over candidates, dominating the conversation while the interviewee

tends to remain passive. The nature of interaction is already determined in the interview test setting; therefore, the participants' roles appear to be fixed in terms of who will lead the conversation and manage the topics, since there is a predetermined script designed to elicit appropriate speaking samples (Kormos, 1999). The routine sequence of one-sided asking and answering the questions are characteristics of this task-type that differ from natural conversation.

Research on traditional face-to face interviews therefore casts doubts on the validity and reliability of that style of interviews. There is a call for making direct oral assessments more reliable and fair through reducing interlocutors' variability and to control their effect. Such arguments have led to the design and development of more valid, reliable, and authentic speaking assessments.

2.2.3 Semi-direct Speaking Assessment as Task-based Assessment and Studies on Task performance

For the purposes of making speaking assessment more valid and reliable, Educational Testing Service (ETS), and the Center for Applied Linguistics (CAL) developed a series of semi-direct format speaking tests. An example of the earlier versions of semi-direct speaking assessment is the Semi-direct Oral Proficiency Interview Test (SOPI), which is a tape-recorded version of OPI (Kuo and Jiang, 1997; Stansfield, 1990) that controls the variability brought about by the interviewer. A few studies have been conducted concerning planning time in semi-direct speaking assessment (Elder, Iwashita, McNamara, 2002; Malanonga, Kenyon, and Carpenter, 2005; Wigglesworth, 1997). With

respect to task topic and intended audience, Lumely and O'Sullivan (2005) investigated the effects of gender on test-takers' responses as well as the effects of gender-oriented tasks on test-takers' scores in tape-mediated speaking assessments. No significant effect was observed regarding the gender of the hypothetical audience, although some differences were reported between male and female test-takers, depending on the topic. One intriguing finding from this study was that test-takers indeed paid heed to the authenticity of task prompts and the concept of interaction. This suggests that more research should be done to shed light on task effects on test-takers' performances from various perspectives, and on their perceptions of tasks.

Semi-direct oral proficiency tests are claimed to have high reliability and validity value since all test-takers perform the same tasks in a similar setting with a fixed amount of time for preparation and for response. This contrasts with direct oral proficiency tests, in which these elements can be flexible. Moreover, advances in computers and related technology have changed the picture of language assessment significantly over the last decade. Technology is used in both education and daily lives, and language assessment is no exception. Speaking assessment is one of the areas where language assessment has benefited most from the application of technology and computer-mediated semi-direct speaking assessment, which will be more common formats in the future. As described in section 1.3.1 above, there are two types of computer-based speaking assessment formats, depending on delivery and scoring: computer-based delivered speaking tests and computer-scored speaking tests. Computer-based speaking assessments have several

advantages and are claimed to be more valid and reliable than traditional face-to -face speaking interviews in several ways.

First of all, test-takers' speech responses elicited from tasks are recorded for immediate or later retrieval, and evaluated by human raters or computers in computer-based speaking assessment. Once computers or related programs have been implemented, the test can be administered anywhere and anytime, and can therefore access a larger population in an identical way, which increases practicality and flexibility (Chapelle & Douglas 2006; Douglas and Hegelheimer 2007; Jamieson 2005; Xi 2010).

In addition to the strengths of practicality and flexibility, computer-based testing may also have high reliability because of its standardized delivery of tasks given to test-takers, which is quite the opposite of face-to-face speaking assessment, in which interlocutor variability might take effect in testing delivery. In light of the scoring process, computer-based assessment could be very efficient, in that digitally stored speech sample files can be sent to certified raters online, allowing them to access the files anywhere and in any order.

Another strength of computer-based speaking assessment is that test takers can engage in more authentic language use tasks that represent real-life situations, since it includes technology-related interactions in the process of communication (e.g., voice mail, and computer presentation graphics (PowerPoint)), which have become predominant these days (Chappelle and Douglas, 2006 ; Douglas and Hegelheimer, 2007). With the growing use of various modes of communication brought about by application of technology,

proficiency should be measured by combining both types of assessment in an effort to uncover meaningful information and a deeper understanding of one's proficiency (van Lier 1989).

One of the characteristics of semi-direct speaking assessments is that the task has increased importance in terms of eliciting test-takers' speaking performances for evaluation since the interviewer is not present and does not interact with the test-taker. This salient trait of computer-based speaking assessment is reflected in its task features designed to elicit candidates' speaking performances. Thus, semi-direct speaking assessments have qualities of task-based assessment (henceforth TBA).

With the need for communicative or authentic language tests, TBA has been given a lot of attention due to the importance of the variety of actual language use it offers. Because language production is a primary concern in real-life domain, measuring communicative language ability is important. TBA considers not only psycholinguistic competencies but also sociolinguistic competencies in various social contexts, where both pragmatic knowledge and language knowledge are needed in order to deal with situations. TBA attempts to measure target language use (TLU; Bachman and Palmer, 1996) in contexts that resemble real-life situations. Thus, TBA has major advantages, which can stimulate both communicatively-oriented language and authentic language use. For these reasons, however, TBA has also been criticized: tasks to elicit conversational language are difficult to create, and authentic language use produced in a particular target language-use domain could lead to restricted interpretations about test-takers'

performances.

Target language use might include interacting with friends and family members in daily life or with clients, customers, and coworkers in a work environment. In case of academic situations, interaction with professors, faculty, and friends could be involved. In addition, such target language use could occur in face-to- face encounters or over the phone.

Table 2.1 shows a summary of characteristics of tasks used in some computer-based speaking tests that are currently administered around the world. In terms of main test purposes, the Versant English test, the Cambridge ESOL Bulats online speaking test, and the TOEIC speaking test are designed to measure speaking ability in business contexts, while the TOEFL test is in academic contexts. Widely used types of tasks in business English tests are answering to given questions, and describing visuals or graphics. In the OPIc speaking test, test-takers are given several situations to role-play. In the TOEIC Speaking test, one task type asks test-takers to listen to a recorded message and leave a voice mail message in response to that message. These tasks seem to be authentic because they deal with a variety of situations that are likely to happen in a real life. However, all the tasks are monologic, and thus there is no interaction involved. Real conversational ability cannot be assessed. This is a major weakness of semi-direct speaking tests, contrary to direct speaking tests, in which speaking ability within a conversational and turn-taking time frame could be assessed.

Table 2.1 Overview of Computer-mediated speaking assessment

Tests	The Versant English Test	The Cambridge ESOL Bulats Online Speaking Test	The TOEIC Speaking test	The ACTFL OPIc Speaking test	The TOEFL iBT Test speaking section
Tasks	6tasks (read-aloud, repetition, short answers, sentence builders, story retelling, and open questions) -63 items in total	5 tasks (interview, read-aloud, presentation about a work-related topic, presentation with a graphic, and communication activity) -11 items in total	6 tasks (read-aloud, describe a picture, respond to questions with and without provided information propose a solution, and express opinion) -11 items in total	Mainly tasks that require responding to questions, asking questions, and proposing solutions in a given situation) -12-17 items in total	6 tasks (2 independent tasks and 4 integrated tasks)
Duration /delivery	17-18 minutes/ over the telephone or on a computer	15 minutes/on a computer	20 minutes/ on a computer	20-30 minutes/ on a computer	13-15 minutes/ on a computer
Topic areas and Situations	Personal information, general interest, and general business environment	Business-related and office situations	Work–related and familiar daily situations	Family, school, work, activities, hobbies, and sports	Familiar topics, campus situations and academic course content
Interactional Relationship	Not interactional/ prompts presented by recorded voice	Not interactional/ prompts presented by recorded voice	Not interactional/ prompts presented by recorded voice	Not interactional/Avatar on the screen with recorded voice	Not interactional/ prompts presented on screen with recorded voice and listening files

2.2.4 Lack of Interaction in Semi-direct Speaking Assessment

Semi-direct speaking assessment, often delivered by computer, uses tasks to elicit test-takers' speech samples that can be evaluated later. Task plays a significantly important role in semi-direct assessments, where language is stimulated in a context that resembles real-life either without the interlocutor or with an imaginary interlocutor. However, as described in Table 2.1 above, tasks used in semi-direct speaking assessment cannot measure communicative competence and consist of constrained forms because of the current limited level of technology. Such constrained tasks used for computer-based semi-direct speaking tests have received criticism for lacking an interaction component. Although they are very practical and efficient in terms of administering the test using those tasks, candidates' actual communicative language abilities cannot be captured and assessed, which is a major threat to the validity of computer-based speaking assessment. This trend across the currently available speaking tests in computer-based language assessment format is seen in the row of interactional relationship in Table 2.1 mentioned above. Such a lack of interaction is also somewhat linked with criticism of the negative washback effect created by computerization; practicing such constrained tasks can neglect a broader range of communicative activities, focusing on limited skills of language (Douglas & Hegelheimer 2007).

To supplement the lack of an interactional component in semi-direct speaking assessment, integrated tasks have been designed. Integrated tasks

have been praised due to two main advantages: test-takers are provided with stimuli on which to base their responses, and tasks that resemble natural communication in the academic context will lead to increased validity (Reed, 1990; Weir, 1993; Wesche, 1987, cited in Lee, 2006).

In a study of prototype integrated tasks for the new TOEFL, Cumming et al. (2004) examined the perceptions of experienced ESL instructors about tasks and the correspondence between their students' performances on those tasks and the performances demonstrated in their classes. The study reported that integrated tasks were perceived as authentic in terms of providing opportunities to elicit abstract ideas and reflect academic situations where students are exposed to lectures or texts and required to explain them orally. However, they were also criticized in that integrated tasks were cognitively and intellectually challenging due to their time constraints, limited provision of visual materials, and the test-takers' lack of familiarity with topics or genres, which may negatively influence students' performances.

Despite the importance of task in semi-direct tests, this area has not been given much attention. Since the new TOEFL iBT test was launched, a small body of literature has investigated the independent and integrated speaking tasks used in the speaking section of the TOEFL iBT, focusing on their effects on test performance (Brown, Iwashita, & McNamara, 2005; Lee, 2006). Lee (2006) and Brown et al. (2005) explored whether test-takers' perform differently or not in the independent and integrated tasks in the aspects of linguistic resources, phonology, fluency, and content. No statistical differences were reported in those features. The researchers claimed that test-

takers' performances were not different on the two task conditions. Recently, Kim (2011) examined the effects of independent and integrated tasks with respect to input (i.e. listening and reading) on test-takers' performances quantitatively and qualitatively. The study found that examinees felt tasks increased in difficulty when a lengthy listening text was integrated. In addition, integrated listening texts more significantly influenced low-level examinees' performances as compared to higher-level examinees, although integrated listening tests did not have much effect on advanced examinees. The study suggests that more research should be conducted to take into account task difficulty and task effects in semi-direct speaking tests, where listening files or reading texts are provided instead of an interlocutor. Overall, empirical studies have been very scant in this regard, which clearly suggests that a greater variety of tasks needs to be designed and developed so that more variety of range of language components can also be captured in the context of semi direct speaking assessment.

2.3 Interactional Competence and Role-plays

2.3.1 Interactional Competence

Interactional competence theory is defined as "a theory of the knowledge that participants bring to and realize in interaction and includes an account of how this knowledge is acquired" (Young 1999:118). The fundamental issue in interactional competence is co-construction that is created by participants in interaction (He and Young 1998; Young 1999). Speaking is jointly constructed

between interlocutors by sharing turns to make communication work. According to interactional competence theory, only local competence exists, and varied types of local competence are acquired through social interaction. That is, co-construction takes place in different communicative speech events in which acquired language knowledge is transferred. With application of the perspective of interactional competence to language learning and testing, it is necessary for learners to be exposed to various social contexts so that they can internalize local competence with practice, and use it in situations of a similar type. This theory pinpoints the importance of identifying interactive oral situations and analyzing them in terms of language functions, language skills, and tasks for more detailed descriptions of the test-taker's ability in each specific interactive oral context.

2.3.2 Studies on Role-plays in Speaking Assessment

Role-play is defined as "Participation in simulated social situations that are intended to throw light upon the role/rule contexts governing 'real' life episodes(p. 224)" by Cohen and Manion (1980, cited in Rosendale (1989)). There are two-types of role-play; dialogic and monologic. Dialogic role-play tasks can stimulate production of communicative competence skills that cannot be easily assessed in traditional face-to-face interview format assessments. Dialogic role-play tasks are less pre-determined than traditional interviews, given that candidates have more opportunities to demonstrate their ability to perform openings and closing, initiate or reject topics, and interrupt. Their interactional and spontaneous nature is one main advantage of role-

plays.

Role-plays have been used as a data generation method in oral proficiency assessment. In fact, role-plays are used as part of subtests in the OPI (Oral Proficiency Interview) or operated to elicit target language use under work-related contexts that simulate the target language use situations (McNamara, 1997). Also, recently developed large-scale semi-direct speaking tests contain monologic role-play tasks to elicit speech samples that represent speaking ability in real life and make an inference about future performance. Despite those benefits, role plays have been given less attention, and thus a very limited number of studies have been conducted.

On the negative side, van Lier (1989) argued that role-play tasks may not be an appropriate measure, since the main ability required to perform role-play activities is acting ability, which is not a particularly important component needed in conversations. However, as opposed to van Lier, Kormos (1999) reported positive evidence for role-play tasks. She examined the data from the Hungarian English Oral Proficiency Exam from the discourse analysis perspective. The study compared a guided role-play task and a non-scripted interview in terms of the degree of dominance and contingency, and the distribution of rights and duties, in order to shed more light on the extent to which communicative competence can be displayed in those two tasks. The findings of the study were that the candidates' ability to perform openings and closings, initiate and reject topics, and interrupt can be assessed better in role-plays than in interviews, thus indicating that role-plays can be a suitable measure to assess communicative competence. This result

was interpreted utilizing the rationale that candidates have more opportunities and rights to manage conversations in role-plays, unlike in interviews where more power is given to the interviewers. Additionally, it was observed that the degree of communicative competence in role-plays could be affected by the presence of explicit instructions that require candidates to take the initiative of topic introduction. That is, more explicit instruction regarding the structure of the interaction prompted candidates to take a more active role.

Another argument in favor of role-plays was explored in Okada (2010). In this recent study conducted to find evidence for the validity of the OPI (Oral Proficiency Interview), the researcher specifically focused on role-play activities. The study investigated what interactional competencies can be manifested by means of a corpus data of dialogic role-play activities from interviews conducted in Japan. The speech samples were analyzed through conversation analysis in order to show that role-plays stimulate co-constructed interaction through turn-taking sequences. The data was analyzed within the framework of interactive footing and interactional competencies. Although asymmetrical or unequal relationships between the interlocutor and the candidate were still found in the interaction of role-plays in the study, since the candidates need to act the given role inside the boundaries the interlocutor and the role-play instructions build, the competencies to construct a coherent interaction by producing turns in a timely and appropriate manner were displayed. The study revealed that what is required to participate in role-plays is the ability to negotiate what to do with the interlocutor through cooperation and acting out a role in a sequentially appropriate manner. This ability is

based on a clear understanding of the previous turn produced by the interlocutor. The researcher argues that such ability cannot be elicited or assessed through interview-led format of direct oral proficiency assessment. The findings from this study suggest that more research needs to be done to shed light on accurate evaluation of communicative competence.

Another study conducted by Halleck (2007) investigated examinees' speech production elicited from two different types of role-plays: the monologic role-play task and dialogic one used in the OPIs. In ACTFL OPIs, interviewees at intermediate and advanced levels are asked to perform role-plays in dialogic situations, while test-takers at superior levels are requested to demonstrate their ability to speak in monologic situations without interaction. By analyzing the discourse between the interviewer and interviewee in the dialogic role-play context, it was found that the interviewer dominated the role-play, providing most of the input. Thus, the interviewee had fewer opportunities to demonstrate language proficiency. This raises a critical question about evaluating the interviewee's interactive competence. Therefore, it was argued that more evidence about proficiency can be drawn from monologic role-play tasks without intervention on the part of the interviewer in terms of gathering more reliable and valid speech samples, which leads to successful and accurate rating.

In a similar vein, Reed and Halleck (1997) investigated the effect of the level of the role-play on the final ratings. The study compared the final ratings of interviews conducted in terms of two conditions where the interviewee took a test with interviewer 1 and interviewer 2. The same interviewee

received higher scores in the interview conducted by interviewer 2, who chose to present a higher level of the role-play that required more monologic discourse. One hypothesized argument for interpreting such findings was that inferences about proficiency were made based on the level of performance, which was a monologic role-play situation in this case. In regard to rating, higher scores were assigned when the performance was good in the monologic role-play, given that speaking by oneself is more difficult because it is artificial, due to the fact that there is no audience. The studies mentioned above were done only in a face-to-face speaking assessment format. Also, it was not explored whether test-takers felt monologic role-plays were more difficult than dialogic role-plays. Investigation into whether such findings are supported is needed requires empirical evidence from a different perspective.

More studies need to be done to shed light on the impact of the choice of monologic versus dialogic role-plays on proficiency ratings with statistical analysis and qualitative analysis. Although dialogic role-plays as an instrument to elicit speech performance could be viewed as problematic, due to the effect of interlocutor variability in the interactive nature of the discourse, the ability to interact still should be investigated in that such skills are essential in real-life communications. In addition, with advances in technology, there are a variety of ways to communicate with people these days. Learners of English will encounter situations where quick responses are required through interaction, as seen in cell phone or mobile communications. There is much more room for investigation for the examination of learners' performances and the nature of language from tasks in various target language

use domains.

Therefore, more tasks need to be developed or investigated in depth to broaden our knowledge of what speaking ability is and how to measure such ability from multiple perspectives and approaches. Also, it is required to make tests more authentic by replicating the target language use domain in task-based assessment. Mobile phone communication is a direct way of communication means and the context is a target language use domain that is essential in modern life. Capturing interactional competence needs to be added to evaluate one's speaking ability accurately in a semi-direct speaking assessment context that is expected to be used more widely in the future. Thus, investigating the feasibility of technically direct speaking tasks over the phone to find out how test-takers perform and perceive interactive tasks as compared to non-interactive tasks would bring insights into how to develop tasks used in semi-direct speaking tests to be more communicative and authentic.

Chapter III

Method

This chapter describes the methodology used to collect and analyze the data for the study. It begins with a description of the participants, the interlocutor, and raters, followed by instruments, data collection procedures, and method of data analysis.

3.1 Participants

Forty Korean learners of English who attended Seoul National University participated in the current study. They were recruited from online postings on the basis of TEPS (Test of English Proficiency developed by Seoul National University) scores. According to the grade description provided by the organizing committee of TEPS, the participants were learners of high- intermediate to advanced proficiency. Their mean score of TEPS was 810, ranging from 701 to 961, and average scores of male (N= 18) and female subjects (N=22) were 822.22 and 797.64, respectively.

Table 3.1 summarizes the background information of the subjects, including the group means for the participants' age and TEPS scores, classified as male and female. The participants belong to a wide range of academic disciplines such as humanities, education, social science, law and science, art and design, and nursing. Among these various areas of

specialization, most of the male subjects were Economics majors while the majority of female subjects majored in humanities. Their ages ranged between 21 and 30, the average age being 23.4 years old. In addition, most of the participants had little experience with studying abroad; 6 out of 40 students reported that they had lived in foreign countries and received at least one year of formal education there. They were compensated for their time for their participation in the study.

Table 3.1 Background information of the participants

Subjects	Male	Female	Total
Number	18	22	40
Age	21~30	21~27	23.4
Major	Economics [a]	Humanities	
No. of students with foreign experiences [b]	3	3	
Average of TEPS scores	825.22	797.64	810

[a] A major area of specialization was Economics for male and Humanities for female subjects.

[b] The number of students who had received at least one year of formal education in English-speaking countries.

In addition to test-takers participating in this study, a native speaker of English from the United States of America served as a conversational partner on the phone acting like a hypothetical friend for each participant in the phone test. There was only one native speaker of English as an interlocutor because the purpose of the current study is to investigate test-taker performance when interaction is involved, not to investigate the interlocutor effect. Thus, the interlocutor effect should be controlled by restricting it to the same person conversing with all test-takers. The native speaker of English is male and

currently a graduate school student at Seoul National University. He was informed of the tasks and trained in order to play his role consistently without much difference among participants.

3.2 Raters

Two raters took part in the current study. Both raters were graduate students majoring in English linguistics at Seoul National University. Before rating all the speech samples of 40 participants, the raters received rater training from the researcher so that they could be well aware of scoring rubrics and would grade the tests in the same way.

3.3 Examiner

The researcher took part in this study as an examiner to administer the four different tasks. The examiner instructed participants to do the two different tests comprised of two tasks and record each individual's speaking performance. However, the examiner did not interact with them.

3.4 Instruments

3.4.1 Preliminary version of the test

One pilot version of the test was conducted in the preliminary study. 24 Korean learners of English and a NS interlocutor, who also took part in the current study, participated in the preliminary study. They were recruited from

online postings on the basis of TEPS (Test of English Proficiency developed by Seoul National University) scores. According to the grade description provided by the organizing committee of TEPS, the participants were learners of high intermediate to advanced proficiency. Their mean TEPS scores were 836.65, ranging from 710 to 967.

In the preliminary study, they performed one non-interactive task and one interactive task. They also were requested to answer a questionnaire asking about test-taking experiences and their preferences. The pilot study found that their performances on the task were not significantly different in terms of the two different types of tasks, and 23 out of 24 participants preferred doing the interactive task.

3.4.2 Non-interactive Speaking Test

The non-interactive test consists of two tasks which are adapted from role-play components of OPIc (Oral Proficiency Interview-computer); the tasks involved proposing solutions to a given problem. This test is used as a non-interactive test for the purposes of comparison with another test in which there is interaction. In addition, since the scores from the non-interactive test were compared with the scores from the interactive test, it was necessary to make the tasks of the two tests parallel. Accordingly, both tests assessed skills that involved reporting a problem and then making suggestions to solve the problem. The participants performed the two tasks with two different situations in the context of leaving a voice mail message to a friend in the non-interactive test. The scores obtained from each task were aggregated to

gain more accurate and reliable scores than would be gained from one single score.

The two situations were presented on PowerPoint slides. All participants were asked to read each situation in the prompt presented on the computer screen and to act each out accordingly by speaking into the microphone connected to the laptop computer for 90 seconds of response time after 30 seconds of preparation time (See Appendix A). Hence, the total amount of time allotted to complete the task was 2 minutes. All the speech performances produced by each participant performing the non-interactive task were recorded via the timer record function of the Audacity sound editing software Version 2.0.0 (Audacity, 2012). Each audio file was given random numbers before being handed over to the raters for the purposes of scoring.

3.4.3 Interactive Speaking Test

The interactive test was devised by the researcher in order to assess test takers' interactive communication competence. Thus, a situation and a role that were given to participants were in parallel with the non-interactive test, but there was interaction in this test. In other words, unlike the non-interactive test, where the participants were asked to speak alone, in the interactive test they needed to talk with a native speaker of English on the phone as if he were their friend. Since a particular target language domain in the non-interactive test is leaving a phone message when the recipient does not answer the phone, the one in the interactive test was devised to take place in a mobile phone conversation context where real interaction takes places (when the recipient

answers the phone), so that more authentic input and output could be elicited. The test consisted of two tasks. Test-takers were required to report a problem, make suggestions, and discuss the problem together to resolve it in each task. The scores from each task were aggregated to gain more accurate and reliable scores than would be gained from one single score.

The instructions and the situation were given via a PowerPoint slide show. The participants read the prompt on the PowerPoint slide and then called their hypothetical friend after 20 seconds of preparation time. The maximum allowable time of communication was 4 minutes (see Appendix B). Additionally, a timer clock was displayed in front of the participants so that they could keep track of time while completing the tasks. For the interactive test, the voice recorder, K9-PRO, was used to record the conversation between each participant and the native speaker of English. This recorder has an ear microphone that can record any conversation when it is placed in either ear with a cellular phone placed against it while the microphone unit is connected to the voice recorder. With this special ear microphone, both sides of the conversation can be recorded clearly on the voice recorder. When each participant called him, the native speaker of English with this ear microphone plugged in his ear hit the recording button. Each participant's phone communication with the interlocutor was recorded at that point. Later, all the speech samples on the recorder were transmitted into the researcher's laptop and saved as MP3 files. Each audio file was given a random number and handed over to the raters for scoring.

3.4.4 Questionnaire for Test-takers

A questionnaire was administered to all participants immediately after they performed the non-interactive and interactive tests, each of which was comprised of two sets of tasks. The questionnaire consisted of three sections: the non-interactive test; the interactive test; and the comparison of two tests. Since the counterbalance design was used to control the order and practice effect for the study, the order of two sections except for the comparison part was different, although the questions that all participants were asked to answer were the same. In other words, participants were divided into two groups; one group that performed the non-interactive test first was asked to answer the questions on the non-interactive test first, while the other group answered the questions about the interactive task first because they did the interactive task first. Thus, two types of questionnaires (A and B) were made (see Appendices C and D). After completing both tests, participants were also asked to discuss their preferences. Most of the questions were related to the participants' test-taking experiences such as test difficulty, appropriateness of test instructions, testing time, preparation time for the test, and their task preferences. Most of the question types used for the questionnaire were 5 band scales (1-5), and also question types utilizing checkbox and paragraph texts were used depending on the characteristic of the questions. The form of the questionnaire was created online by using Google Docs and printed out in hard copy.

3.4.5 Questionnaire for Raters

A questionnaire was administered to raters to investigate how the raters perceived the two different types of tests and to provide valuable feedback on scoring criteria. The questionnaire consisted of two sections: prior to rating and post-rating. (see Appendix E). Raters were asked to answer prior-to-rating questions before they were given all the speech samples, while answers to post-rating questions were collected after they completed scoring.

3.4.6 Scoring Rubric

Each participant's speech samples collected from the non-interactive and interactive tasks were scored by two raters based on the scoring rubric. Two analytic scoring rubrics were designed specifically for the present study (refer to Appendix G) and were used for the non-interactive test and interactive test. An analytic scale was adapted from other scales used for assessing speaking skills, such as the ones used for the TOEIC Speaking and IELTS tests. It also took into consideration the rating scale developed by Nakatsuhara (2007) that was designed for assessing English speaking in group oral activities. Another rubric taken into consideration was an analytic scoring rubric (cited in H.J Kim, 2011) used to evaluate students' responses as part of placement tests in the ESL program at Columbia University. This scoring rubric is based on speaking ability as defined by Purpura (2004), including phonological control, control of grammatical forms, control of conversational structure, meaningfulness, and control of pragmatic meanings.

The analytic scale was used because it is finer-grained than the holistic scale, consisting of five bands for five separate criteria, which enables us to discern any possible differences in the speaking performances of participants depending on the conditions: whether interaction is involved or not. In addition, somewhat different scoring criteria for different task types can be applied if analytic scoring is used. The criteria for assessing speaking skills in the non-interactive test are task completion and discourse management, pronunciation and intonation, control of grammatical forms and vocabulary, fluency, and control of pragmatic meanings while the ones for assessing speaking skills in the interactive test are pronunciation and intonation, control of grammatical forms and vocabulary, fluency, control of pragmatic meaning, and interactional competence. There are five levels of performance (1 to 5) for 5 criteria. So, the maximum possible aggregated scores for all five criteria combined are 25, and the analytic scores divided by 5 criteria were reported. It should be noted that while three criteria (pronunciation and intonation, control of grammatical forms and vocabulary, and fluency) are the same in that they are concerned with language, the last two criterion of the rating scale for each test are different in the sense that they are concerned with pragmatic and sociolinguistic knowledge specifically for the task contexts and the degree of task completion required to perform speaking in each task. That is, success of conveying messages and interacting with the person in phone communication contexts is measured respectively in each task by using the different criteria. Such criteria were deemed important to be included as part of the evaluation scale since knowledge regarding the context and register is

needed, and thus it should be reflected in how candidates perform and complete the task.

3.5 Data Collection Procedure

Participants were divided into two groups; group A and group B. To avoid practice effect, the order of the tests was counterbalanced. In other words, half of the participants in Group A performed the non-interactive test composed of two voice mail tasks first, while remaining participants in Group B performed the interactive test comprised of two phone tasks first. Testing sessions were held in quiet rooms using a laptop computer and a cellular phone, and took a total of four days. Each participant and the examiner were in a room while the interlocutor receiving phone calls from the participants was in another room.

The participants were greeted and instructed about what to do by the examiner. After signing the informed consent form (refer to Appendix H), participants began to perform the two tests, sitting with a laptop computer facing them on the desk. A timer was placed in front of them. Between the two tests and after completing the last one, they were asked to fill out the questionnaire that consisted of three sections —section 1 for the first test type, section 2 for the second test type, and section 3 for the comparison on the two tests. The description of how data were collected is summarized in
Figure 3.1.

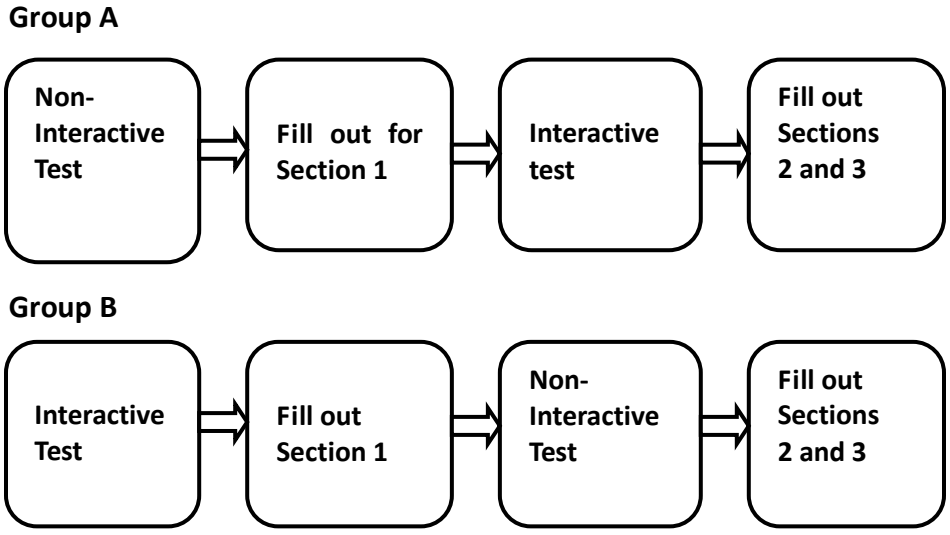


Figure 3.1 Description of the test administration procedure

3.6 Method of Analysis

All scores submitted by the two raters for each test were first keyed into a Microsoft Excel 2007 spreadsheet and then transferred to IBM SPSS Statistics Version 20.0. This software was used to obtain the descriptive statistics for the analytic scores, correlation coefficient among task, test scores and other criterion measures, and reliability coefficients for ratings.

To answer the first research question, in terms of reliability, the degree of agreement between the two raters (inter-rater reliability) was investigated from various perspectives, such as from the participants' total score, across the two tasks in each test, and across the five rating scales by calculating Kappa coefficients. The agreement indexes were also computed for the adjudicated scores. All these reliability estimate measures served as evidence to investigate the construct validity of the scores from the speaking test

developed for the purposes of this study. In terms of examining validity, the relationship between the scores from the non-interactive test, the interactive test and TEPS as a standardized measurement of language proficiency, the Spearman rank-order correlation coefficient, (Pearson product-moment correlations) were computed.

To answer the second research question, the speaking performances from the non-interactive and the interactive tests, when participants were asked to speak alone and when participants were asked to talk with the interlocutor on the phone, were compared to observe whether participants performed differently in the two tests according to their composite scores, total speech length, and conversational turns in the case of the interactive task. In order to address the third and fourth research questions, qualitative analysis was conducted. The participants' responses on the questionnaire were collected and tallied. The rating scale (1 "strongly disagree" to 5 "strongly agree") were used.

Chapter IV

Results

This chapter presents the results of the statistical analysis of raw scores and rater-reliability. It begins with the descriptive statistics of the NI test and Interactive test, followed by the rater-reliability, construct, and concurrent validities of the tests scores. Next, it examines how long the test-takers perform under the two different test conditions and how many conversation turns take place in the interactive task with provision of several samples. It will end with provision of the test-takers' responses, the raters' feedback, and the interlocutor's feedback regarding the tests, scoring rubric, and test-takers' performances.

4.1 Descriptive Statistics for Non-interactive and Interactive Tests

Table 4.1 shows the descriptive statistics of the performances in two speaking tests, namely are the non-interactive test and the interactive test. Each test consists of two tasks. To examine the trends of the participants' performances across all of the criteria, the means and standard deviations of the scores on each criterion of the two tests are also provided in Table 4.2 and 4.3.

As Table 4.2 shows, the mean scores on the control of pragmatic meaning criterion in both tasks in the non-interactive test are the highest among all of the criteria. This indicates that the degree to which the

participants displayed their pragmatic knowledge was higher compared to the other areas of criteria. In terms of variability in performance, there was a greater variability in the control of grammatical forms and vocabulary criteria among the participants, given that the standard deviations for the criterion were the largest among all of the criteria in the non-interactive test. As for the interactive test, the mean scores on the control of pragmatic meaning criterion in both tasks were higher than those on the other criteria.

Regardless of test-type, the lowest mean scores were found on the control of grammatical forms and vocabulary. This indicates that the degree to which the test-takers have mastered grammar and vocabulary was lower compared to other criteria areas. As shown in these tables, it was found that the composite scores of the participants on the non-interactive test were higher than those on the interactive task, which could be interpreted as the overall performance of the participants being better in the non-interactive test than the interactive test.

Table 4.1 Raw scores of the non-interactive and interactive tests

Tests		N	Mean	SD
NI	Task 1	40	18.36	3.88
	Task 2	40	18.18	3.74
I	Task 1	40	17.64	3.67
	Task 2	40	17.89	3.72

Table 4.2 Descriptive statistics for scores from Non-interactive tasks according to criterion

NI	Task 1(Tour guide)		Task 2 (Musical ticket)	
	Mean	SD	Mean	SD
T/D	3.84	.83	3.69	.89
P/I	3.64	.94	3.52	.91
F	3.64	.92	3.74	.79
G/V	3.40	.97	3.39	.99
P/M	3.85	.89	3.84	.78
Composite	18.36	3.88	18.18	3.74

Table 4.3 Descriptive statistics for scores from Interactive tasks according to criterion

I	Task 1(Trip)		Task 2 (Birthday party)	
	Mean	SD	Mean	SD
P/I	3.44	.83	3.38	.99
F	3.45	.94	3.50	.83
G/V	3.31	.92	3.35	.98
PM	3.80	.97	3.81	.84
IC	3.64	.89	3.85	.85
Composite	17.64	3.67	17.89	3.72

Table 4.4 Descriptive statistics for the average scores from the Non-interactive and interactive tasks.

	Mean	SD
NI	18.27	3.70
I	17.76	3.57

4.2 Reliability Measures

This section reports the inter- and intra-rater reliability of test scores the correlations among test scores and TEPS, and the correlations between test tasks. Reliability measures should be dealt with to provide evidence of concurrent and test validities of the speaking scores along with rater reliability

for all of the obtained scores.

4.2.1 Inter-rater Reliability

The Spearman rank-order correlation coefficient was computed to examine the inter-rater reliability. Since speaking performances were scored based on an analytic scoring rubric, correlations between rater 1 and rater 2 were calculated for the scores in each analytic criterion and the composite scores. Later, score agreement rates between the two raters were calculated in terms of perfect agreement, adjacent agreement, and non-adjacent agreement indices, as well as the percentage, in order to assess the proportion of agreement between raters regarding the scores given to each criterion in each test. The Kappa coefficients and Spearman-Brown predicted reliability coefficients were also computed as further evidence of inter-rater reliability.

Table 4.5 shows the Spearman rank-order correlation coefficient results between the two raters in scores given to each rating criterion in each test. Overall, a significantly high correlation between the two raters was found in the NI test (. 84 for task 1 and .75 for task 2) and I test (.90 for task 1 and .86 for task 2). With regard to each criterion, the correlation value range was from .49 to .82 for task 1 and 2 respectively in the NI test and from .61 to .84 for task 1 and 2 respectively in the Interactive test.

Under the NI condition, lower correlation coefficients as compared to other criteria were found in the criterion of PM (control of pragmatic meanings) for both tasks, although scores in T/D (task completion and discourse management) criterion had the lowest correlation coefficient for

task 2. Those scores indicate that the two raters had a greater degree of disagreement when scoring in those criteria. This could also be interpreted as indicating that test-takers' performances in those criteria was particularly difficult to score. However, those lower coefficient values increased when the third rater's adjudication was taken into account.

Under the condition where interaction is involved, the lowest correlation coefficients were calculated in the criterion of IC (interactive communication) for both tasks. This could be an indication that interactive communication was particularly difficult for both raters to measure, which lead to a greater degree of disagreement. This could also be interpreted as being because the IC scoring rubric might have caused confusion for both raters due to a vague descriptor in the rubric. The correlation values did not increase much when the third rater adjudicated on the score.

Table 4.5 Spearman rank-order correlation coefficients between raters in the non-interactive (a) and interactive (b) tests according to each criterion

a)

SC	NI Condition			
	Task1		Task2	
	Spearman Rho	Pearson Correlation	Spearman Rho	Pearson Correlation
T/D	.73	.74	.49(.70)	.48(.66)
P/I	.82	.82	.80	.79
F	.75	.74	.68(.77)	.69(.77)
G/V	.69(.86)	.69(.85)	.70	.73
PM	.62(.77)	.62 (.77)	.51(.60)	.51(.60)
Total	.84	.85	.75	.78

Notes: All significant at .01 level (2-tailed)

( )=when considering the third rater's adjudication

SC= Scoring criteria

b)

SC	I Condition			
	Task1		Task 2	
	Spearman Rho	Pearson Correlation	Spearman Rho	Pearson Correlation
P/I	.82	.83	.77	.78
F	.80	.79	.80	.75
G/V	.77	.78	.82	.81
PM	.84	.80	.68	.70
IC	.63 (.68)	.64 (.68)	.61(.61)	.64(.64)
Total	.90	.90	.86	.87

Notes: All significant at .01 level (2-tailed)

( )=when considering the third rater's adjudication

SC= Scoring criteria

Table 4.6 presents agreement indices between raters, along with the Kappa coefficients and Spearman-Brown reliability coefficients in the NI test. Perfect agreement means that there is no score discrepancy since the two raters gave the same score. Adjacent agreement represents the agreement rate of scores differing by only 1 band. Non-adjacent agreement represents the proportion of scores differing by 2 bands or more. As shown in Table 4.6, perfect + adjacent agreement rates range from 77.5% to 100% for the two tasks in the non-interactive test. Although a relatively high degree of agreement was found between the two raters, non-adjacent agreement cases were also observed, which accounts for the rate ranging from 5% to 22.5%.

The Kappa coefficient claims to take the agreement rates occurring by chance into account. Thus, the Kappa coefficients were also computed in order to provide a better measure of the inter-rater agreement. The results show that there was a slight to fair level of agreement (Landis & Koch, 1977, cited in Sim & Wright, 2005). When the third rater adjudicated on the non-

adjacent scores, the Kappa coefficient value increased slightly or stayed the same while one in the P/I (pronunciation and intonation) criterion decreased.

Spearman-Brown reliability coefficients were also provided by the Spearman-Brown prediction formula (Spearman, 1910). This formula is used to predict the reliability when the test length or the number of test items increases; these coefficient values are expected to increase when more items are added to the test. The results showed that the predicted reliability increased as presented in table 4.6.

Table 4.6 Score agreement rates and Kappa coefficients between raters in the non-interactive test

NI	SC	Perfect Agreement	Adjacent Agreement	Perfect + Adjacent	Non-Adjacent Agreement	Kappa	α
		Rate	Rate	Rate	Rate		
Task 1	T/D	.58	.43	1.00	-	.38	.84
	P/I	.33	.48	.80	.20	.14 (.12)	.90
	F	.48	.48	.95	.05	.30 (.30)	.85
	G/V	.33	.45	.78	.22	.15 (.17)	.81
	PM	.45	47.5	.92	.08	.25 (.25)	.76
Task 2	T/D	.33	.53	85	.15	.10 (.16)	.65
	P/I	.38	.45	.82	.18	.17 (.20)	.88
	F	.53	.43	.95	.05	.32 (.36)	.80
	G/V	.30	.58	.88	.12	.15 (.80)	.82
	PM	.40	.53	.92	.08	.15 (.15)	.67

Notes: ( )=when considering the third rater's adjudication.

Adjacent Agreement= the agreement rate of scores differing by +/- 1 point.

SC=Scoring criteria

It is noteworthy that in spite of the high correlation coefficient value between the rater's scores, only a slight to fair level of agreement (from .10 to .38) was obtained. This implies that although the raters scored similarly in rank ordering the speaking performances produced by the test-takers, raters' severity levels were different; one rater was stricter while the other rater was more lenient in terms of scoring.

Table 4.7 Score agreement rates and Kappa coefficients between raters in the interactive test

I	SC	Perfect Agreement	Adjacent Agreement	Perfect + Adjacent	Non-Adjacent Agreement	Kappa	α
		Rate	Rate	Rate	Rate		
Task 1	P/I	.38	.60	.97	.03	.12 (.12)	.90
	F	.50	.50	1.00	-	.32	.88
	G/V	.50	.48	.97	.03	.29 (.29)	.87
	PM	.60	.40	1.00	-	.34	.91
	IC	.30	.68	.97	.03	.06 (.06)	.77
Task 2	P/I	.33	.55	.87	.13	.13 (.12)	.87
	F	.53	.45	.97	.03	.33 (.33)	.88
	G/V	.35	.65	1.00	-	.16	.90
	PM	.43	.53	.95	.05	.24 (.22)	.80
	IC	.35	.65	1.00	-	.09	.75

Notes: ( )=when considering the third rater's adjudication
Adjacent Agreement=the agreement rate of scores differing by +/- 1 point.
SC=Scoring criteria

Table 4.7 displays the agreement rates between the raters' scores and the Kappa coefficient, along with the Spearman-Brown predicted reliability in the

interactive test. The results showed that the predicted reliability increased. As compared to the NI test, fewer non-adjacent agreement cases were found in the Interactive test, with the rate of perfect agreement + adjacent agreement ranging from 87.5% to 100%. The interactive test also obtained slightly lower Kappa coefficient values, with a slight to fair level of agreement (from .06 to .34). IC was the criterion with the lowest Kappa value for each task in the interactive test, having the lowest rate of perfect agreement (30%). These results are in line with the lower Spearman rank-order coefficients in scores given to the criterion of IC. The pattern of high correlation but a lower level of agreement between the raters' scores found in the NI test was observed in the Interactive test, which demonstrates raters' different levels of severity in scoring.

4.2.2. Correlations among Test scores and the Criterion

The relationship among the test scores should be examined in order to validate the tests designed for the present study. Since the participants were recruited on the basis of their TEPS scores for this study, how much their English proficiency measured by TEPS correlates with their performance on the two tasks in the non-interactive and interactive tests was examined using Spearman rank-order correlation coefficients. Additionally, their performances on each test when scored by rater 1 only and rater 2 only were also calculated for further examination of correlation. As shown in Table 4.8, strong positive correlations among the non-interactive and interactive tests and the TEPS were obtained with coefficients of .73 and .70, respectively. Coefficients

were .72 for the non-interactive test and .71 for the interactive test when the performances were scored by only rater 1 and only rater 2, respectively. Thus, the overall range of coefficients was from .70 to .73. Also, in terms of relationship between the two tasks, the scores of the two tasks in the NI and I tests were highly correlated, with a range of .83 to .86, respectively. Such high coefficients between different test scores indicates a strong positive relationship among items and tests, which gives rise to evidence for item discrimination and validity of the tests designed for the present study.

Table 4.8 the Spearman rank-order correlation coefficients between tests

Test Scores	NI-Test					Interactive-Test					TEPS
	T1	T2	Total	R1	R2	T1	T2	Total	R1	R2	
NI-T1	1									.	
NI-T2	.83	1									
T-NI	.95	.94	1								
R1-NI	.91	.93	.97	1	.						
R2-NI	.92	.88	.93	.84	1						
I-T1	.81	.83	.85	.87	.78	1					
I-T2	.74	.75	.77	.78	.72	.86	1				
T-I	.80	.83	.85	.86	.79	.96	.95	1			
R1-I	.80	.83	.85	.86	.78	.94	.94	.98	1		
R2-I	.76	.78	.80	.81	.75	.94	.94	.91	.91	1	
TEPS	.73	.70	.73	.72	.71	.69	.67	.70	.70	.70	1

Notes: All significant at .01 level (2-tailed)

4.3 Analysis of Test-takers Language Samples

It was deemed necessary to examine if test-takers performed equally well on both conditions or outperformed in one condition over the other to answer the second research question. The following Figure 4.1 illustrates examinees'

performances on the non-interactive and interactive conditions. As presented in the figure, there are some cases where individual differences were captured between the two test conditions. Some students performed better on the non-interactive test condition while others received a higher score in the interactive test condition. Several who did equally well on both conditions were also found. Those three patterns are categorized with the test-takers' number and the total number of tokens in Table 4.9.

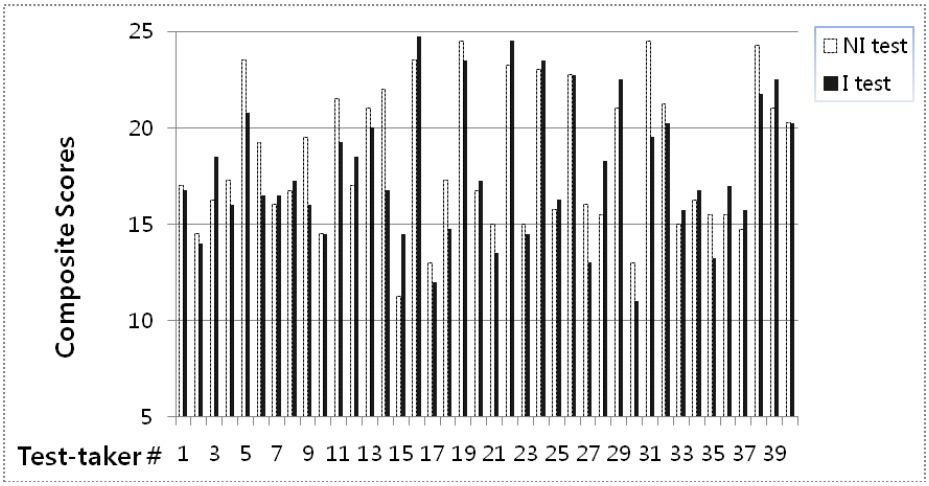


Figure 4.1 Composite scores for the non-interactive test and the interactive test

Table 4.9 Performance comparison in the non-interactive test and the interactive test with the total number of tokens

	NI > I = performed better in the NI test	NI < I =performed better in the I test	NI=I =the same performance
The Total Number of Tokens	20	17	3

Based on the pilot study, 90 seconds for the non-interactive task and 4 minutes for the interactive task were allocated as an appropriate length of time to complete each. It is no surprise that some individual differences were involved in terms of speed or rate of performance during those limitations in speaking performance time. In order to illustrate how each test-taker performed, the length of time each test-taker took to complete the two non-interactive tasks is provided in Table 4.9. The number of conversational turns from the two interactive tasks was counted to find out how many turns were shared between the two conversational partners during the total conversation time in the interactive phone communication task. One interesting finding is that performance-time length does not always directly correlate with the quality of performance. For example, test-taker 23 spoke more than the given test time in one task and used almost the whole time to speak in the other task on the non-interactive condition. The same test-taker spoke with the interlocutor for 4 minutes and 28 seconds (longer than the given time) in interactive task 1 and talked for 3 minutes and 35 seconds in task 2. But only 20 and 16 conversational turns occurred respectively in the two tasks. In fact, the test-taker spoke with frequent pauses, and conversational breakdown occurred, resulting in a longer time length. Conversely, test-taker 22 finished both non-interactive tasks within forty seconds. As for the interactive condition, 36 and 31 conversational turns occurred in each task. That is, the test-taker had a conversation with the interlocutor in a spontaneous and fluent manner.

Table 4.10 Total time in the NI test and total time and the number of conversational turns in the I test for each test-taker

Subject	NI Task 1	NI Task 2	I Task 1		I Task 2	
	Total time (Second)	Total time (Second)	Total time	# of turns	Total time	# of turns
1	74	67	2’38”	29	2’54”	24
2	58	56	2’49”	30	2’23”	22
3	73	72	2’59”	27	2’47”	21
4	90	52	2’59”	31	3’18”	28
5	68	50	2’00”	26	2’41”	25
6	75	74	3’57”	28	1’45”	18
7	79	56	3’21”	34	2’41”	22
8	90	90	3’09”	28	4’17”	30
9	44	62	2’42”	25	2’40”	21
10	88	90	3’38”	29	3’07”	28
11	66	60	2’53”	30	3’09”	28
12	85	90	2’32”	21	3’14”	26
13	88	67	3’36”	30	4’19”	30
14	80	75	3’35”	30	3’37”	31
15	88	90	2’24”	24	2’56”	24
16	73	75	2’32”	26	2’38”	24
17	65	62	2’46”	24	2’48”	26
18	82	53	3’03”	23	3’51”	29
19	74	65	3’29”	27	2’45”	25
20	88	89	4’51”	31	4’00”	29
21	83	67	4’16”	31	4’08”	32
22	40	33	3’12”	36	2’40”	31
23	90	88	4’28”	20	3’35”	16
24	90	87	3’29”	20	3’17”	25
25	90	69	3’23”	20	3’32”	19
26	87	83	4’50”	38	3’21”	25
27	78	67	3’50”	26	4’16”	29
28	38	34	3’38”	25	2’40”	25
29	44	35	2’53”	19	2’34”	19
30	76	42	3’48”	20	3’15”	20
31	80	72	2’48”	19	2’52”	25
32	47	48	2’57”	24	4’07”	29
33	62	50	3’03”	23	3’26”	26
34	38	37	2’54”	23	2’47”	22
35	50	34	3’34”	23	3’42”	24
36	70	79	3’08”	27	3’02”	27
37	59	90	2’59”	18	4’00”	28
38	68	81	5’19”	39	4’00”	23
39	43	59	4’00”	36	2’51”	21
40	54	44	2’46”	26	3’11”	24

Additionally, some transcripts of the recorded responses of several examiners are presented to illustrate the language features they produced. The following is the recorded responses:

a) Task 1 for the non-interactive condition

Test-taker #17

Hey, Alex. I'm Jiwon. I promised you to introduce Korea next week. but, I suddenly I have a fair on day. So, I can't play around you with one day. How about you going out with my friend? My friend is, my friend wants to see you, and she is very wondering about you. So, I want you and my friend to be very best friends. So, one day, can you play around with my friend? If you OK, call me later.

Test-taker #37

Hello, Alex. It's me, Gordon. I wanted to call you to explain the situation, but since you're not answering the phone, I'm just leaving this message. So, the problem is that it seems like I can't take you to places in Wednesday because I have this big job fair happening on that Wednesday. So, during that day, I have to be concentrating solely on that one. I may not be able to help you that day. So, I've been thinking of some ways to solve the situation. I think I can introduce you to one of my friends who can help me doing this thing instead of me. Or, I can. you know, arrange a meeting between you and some other foreign students who are in the kind of similar situation. I think these are two ways to help you. But, I'm not sure what you're gonna prefer. So, if you check this message, Please call me as soon as possible. Thank you.

b) Task 1 for the interactive condition

Test-taker #30

I: Hello

S: Hello

I: Hey, what's up?

S: We have a travel plan next weekend.

I: Right, we were going to Jeju-Island.

S: But, this Saturday I have a team project meeting, which is so important in my grade.

I: OK, that's too bad.

S: I do not make a trip this weekend

I: Are you going to be busy all weekend?

…

Test-taker #26

I:Hello?

S: Hello? This is Jin.

I: Oh, hey, What's up? How are you?

S: Pretty good. Long time no see man

I: Ye, it's been a while

S: Yeh, I'm calling you to let you know something because I have a big problem with our trip to Jeju

I: Oh, what's wrong?

S: You know we are supposed to go to Jeju Isalnd from this Friday to Sunday

I: Ye, that's right.

S: But, unfortunately my crazy professor made a team project meeting on this

Saturday.

I: That's sucks.

S: I know. I want to skip it, but I have to go there because of my grade. If I get a F grade, I can't graduate this semester. I really want to go to Jeju Isalnd, but I have to.

I: I know. Jeju is much better than a team project. but, you can't do anything about it. OK, what should we do instead?

…

Transcripts of the recorded responses for the interactive tasks consist of the beginning parts only. The above language samples in the extracts show some noticeable features such as grammatically incomplete sentences, incorrect use of parts of speech, and inappropriate collocations depending on test-takers' proficiency levels. The noticeable differences between subject 17 and 37 for the non-interactive task are the amount of speech they produced, the complexity of sentences, and a detailed explanation to make an excuse or to propose a solution. For the interactive task, subject 26 had a conversation in a more active and appropriate manner by responding to the interlocutor's questions compared to subject 30. Subject 30 did not respond to the interlocutor's reactions appropriately, focusing only on what was required to complete the task, rather than having a conversation by interacting with the interlocutor. This pattern seems consistent among the subjects of lower proficiency levels, although not all of their transcripts of the recorded speech samples were analyzed for this study. The majority of the test-takers appear to have struggled with responding to the interlocutor naturally and spontaneously,

although they accomplished the required tasks given in the instructions fairly well.

4.4 Test-takers' Perceptions

In order to supplement the quantitative findings, questionnaires were distributed to test-takers. Responses to the questions of post-test questionnaires (the non-interactive and interactive test) were collected and tallied. The results of the tally are presented in Table 4.12, and also shown graphically in Appendix I.

As can be shown in Table 4.12, patterns of responses to most items were quite similar. In terms of nervousness during test performance, the majority of the participants agreed on the statement that they felt nervous while doing the task (65% for NI, 82.5% for I above scale 4). This seems to provide strong evidence that test takers tend to get nervous when they take a test regardless of task format.

When considering self-assessment of performance in the two tests, although more than half of the participants (65% for NI) and almost half of the participants (47.5% for I) felt that they had performed poorly in both tests, it seems that they felt that they performed better in the interactive test (15%) than the non-interactive task (7.5%) (52.5% and 35% in chose of above 3 on the Likert scale for I and NI, respectively).

In response to testing time and instructions, the majority of the participants (77.5% for NI and 75% for I) agreed that they understood what they were supposed to do. They disagreed on the statement that the time

allowed for the task was too short (17.5 %for NI and 20% for I) while there was some agreement on the statement that more preparation time is needed (37% for NI and 47.5% for I); it is predictable that most participants believed they would have performed better if they had been given more time to prepare. Some participants, when asked the most difficult part of the non-interactive test, wrote that it was difficult to think about more than two suggestions during the given preparation time in the paragraph-type question. Thus, with more preparation, they felt they could have come up with more suggestions in a coherent and detailed manner. However, this tendency was slightly stronger in the interactive test context when they talked to a native speaker on the phone, and such responses were in line with some of the participants' comments that "although there was preparation time given, I could not use all I had prepared because I needed to react to the speaker instantly, but I could not anticipate what the speaker would say."

As for perception of difficulty, about a third of the participants (30%) felt that the non-interactive task was difficult, whereas 12.5% of the participants felt that way in the interactive task. Those who disagreed on the difficulty of the test accounted for 55% and 45% of the response for the non-interactive and interactive tasks, respectively. In general, the participants seem to have felt the level of difficulty moderate, although they perceived that the non-interactive test was slightly more difficult than the interactive test.

Table 4.11 Overview of the questionnaire responses I

Item	Task	Strongly disagree 1 %	2 %	3 %	4 %	Strongly agree 5 %
Nervousness	NI	0	7.5	22.5	45	25
	I	5	5	27.5	42.5	40
Believe performed well	NI	20	45	27.5	7.5	0
	I	17.5	30	37.5	15	0
More task time needed	NI	20	27.5	35	15	2.5
	I	17.5	37.5	25	17.5	2.5
More preparation time needed	NI	2.5	37.5	22.5	27.5	10
	I	12.5	25	15	40	7.5
Understood instructions	NI	2.5	2.5	17.5	60	17.5
	I	0	5	20	60	15
Task too difficult	NI	17.5	27.5	30	27.5	2.5
	I	15	30	42.5	10	2.5
Task interesting	NI	2.5	5	35	35	22.5
	I	0	2.5	7.5	65	25
Task realistic	NI	0	10	17.5	45	27.5
	I	0	2.5	2.5	40	55
Demonstrate my ability	NI	0	10	17.5	45	27.5
	I	0	5	20	55	20

In addition, almost all participants found the interactive test realistic and authentic with 95% agreement. It seems as if the fact that the interactive test is based on a phone conversation context, which was easier to relate to their daily life situations, likely contributed to such a high percentage of the participants agreeing on the statement. This finding was also supported by the

participants' written comments. While 72.5% agreed on the statement regarding the non-interactive test, some participants commented that "*I personally don't leave phone messages. I have never done this before, so I didn't know what to say. It was awkward*," and "*it was very difficult because I have no idea of what a NS would do in this context of leaving a voice message*."

When considering adequate opportunities to demonstrate the ability to speak English, three quarters of the participants believed that their ability could be demonstrated through the interactive test, with 72.5% agreeing on the statement regarding the non-interactive test.

Several questions were additionally asked to the test-takers in the interactive test section in order to explore their perceptions of interaction with the native speaker on the phone. The participants' responses are summarized in Table 4.12. They positively responded to the statement that they understood what the person on the phone was saying well, with 82.5% agreement rate. In response to the question about the interlocutor, 95% of the participants did not believe that they would have obtained better scores if they had talked with a different interlocutor, with none (0%) disagreeing with that statement, which shows that the participants did not have any complaints or difficulty in talking with the native speaker participating in the study. It is noteworthy that they did not believe there would be much difference if they had talked with a different partner.

In light of the concerns of preferring face-to-face communication, 57.5 % of the participants felt that they would have performed better if they had

talked with the person face–to-face while 32.5% of them disagreed with the idea. The tendency to prefer talking face-to-face was supported by several participants' comments. They stated that they found it difficult to converse with a native speaker on the phone rather than interacting in person because they could not read the partner' gestures or facial expressions**.**

Table 4.12 Overview of the questionnaire responses II

Item	Task	Strongly disagree 1 %	2 %	3 %	4 %	Strongly agree 5 %
Understood Interlocutor	I	0	7.5	10	50	32.5
Different interlocutor =>better score	I	45	50	5	0	0
Face-to-face =>better score	I	12.5	20	10	32.5	25

The last section of the questionnaire consisted of questions asking about participants' preferences between the two tasks. When requested to select which test they felt to be more difficult and were nervous about taking, 21 of the participants (52.5%) chose the non-interactive test to be more difficult and causing greater anxiety, with 19 participants (47.5%) responding that they felt that way when they did the interactive task. Interestingly, the number of such respondents in terms of the level of difficulty corresponded to the number of respondents in terms of the level of anxiety. Whilst little difference between the two test types was found in terms of the level of anxiety and difficulty, a

few participants mentioned in their written comments that they felt more nervous when they could not anticipate what the speaker would say and when they made the partner wait while looking for appropriate words during the ongoing conversation, which was the most difficult aspect of the interactive test. It would have been much more difficult for the participants who do not have many opportunities to talk on the phone in English, not to mention the natural anxiety of being assessed in a testing context.

In response to which task they believed would show a more accurate picture of their ability to speak English in real life situations, 38 participants, which accounts for 97.5% of the respondents, selected the interactive test. When it came to the question of which test they would prefer to do for assessment, 39 participants accounting for 95% chose the interactive task over the non-interactive task. When asked to explain why they preferred the interactive test type, many reasoned that they felt talking with a person was more natural, relaxed, and comfortable than talking to the microphone alone. Also, some participants commented in their written comments that although they misunderstood some parts of the task that they were required to do or made some mistake in the beginning, they felt that they could make up for their mistake by interacting with the person in that the respondent reminded them of what they needed to do in order to complete the task.

Although there were individual differences regarding how they perceived the two test types in terms of difficulty and anxiety, 38 out of 40 participants revealed a clear preference for the test with interaction. They stated this preference in their written comments as follows:

1) *"The phone conversation test appears to be more meaningful assessment than leaving a voice message test. It is more like a real-life situation, since it involves listening as well, which can assess listening ability along with speaking ability."*
2) *"The phone communication task is more reflective of real communication."*
3) *"I think experiencing this test of talking with a native speaker of English helps me overcome foreign language anxiety. "*
4) *"The talking on the phone task gives a better display of real-life situations and impromptu speech"*
5) *"We solve problems instantly by interacting with others when encountering problems in a real-situation, so the phone conversation task reflects more reality in the sense that such types of speaking will occur more often in daily life or work contexts."*

On the other hand, the two participants who preferred the non-interactive test stated that "It is easier to prepare," and "it is similar to typical tasks used in other speaking proficiency tests, so it cause less anxiety, and I think it will be easier to get good scores with some effort."

4.5 Feedback

4.5.1 Rater's feedback

The raters were asked to provide prior and post-rating feedback by responding to a short questionnaire about the non-interactive and interactive tasks, scoring rubric, and test-takers' performances (refer to Appendix E). In terms of the overall opinion about the tests, the raters responded positively about the interactive test, given that tasks used in the semi-direct speaking

assessment lacked interaction.

Rater 2 commented about the difficulty of the non-interactive task in that the situation where a caller leaves a voice mail is rare in the Korean context. In contrast to western culture, Koreans rarely leave messages but hang up the phone when there is no answer. For this reason, he argued that the test-takers felt awkward and found it difficult to do the task. In the same vein, those who have had no experience of being exposed to such an environment could have felt intimidated to perform the non-interactive test because of confusion about what to measure in the task completion and discourse management criterion and the criterion of control of pragmatic meanings, which are closely related to acquisition of sociolinguistic competence.

When requested to comment about interesting things or patterns they observed, rater 1 wrote that she observed that the interlocutor in the interactive test sounded nicer to female test-takers than male test-takers. Rater 2 mentioned that he found that overall, the interlocutor sounded friendlier to the test-takers who initiated the conversation with a friendly tone and were willing to negotiate. Additionally, willingness to interact well might have been affected by personal variables such as personality and ability to manage test anxiety. In terms of test-takers' performances, Rater 1 stated that there were a few cases where those who performed well in the non-interactive test did not do well in the interactive test since they were not receptive to the interlocutor, and she had to assign them fewer points in the interactional communication criterion.

In regard to the scoring rubric, both raters agreed that assigning scores in

the criterion of control of pragmatic meanings and the interactive communication criterion seemed to be the most difficult aspect since there were some overlapping points in the two criteria, which made them hard to differentiate from each other. They suggested components in the two criteria be differentiated clearly and explicitly for more accurate evaluation.

4.5.2 Interlocutor's feedback

The third attempt to supplement quantitative findings was to gather feedback from the interlocutor. He provided feedback by responding to a short open questionnaire about the interactive test, the experience in interacting with a total of 40 test-takers, and their performances (refer to Appendix F).

One of the interesting observations he mentioned was that the majority of the test-takers did not introduce themselves and were direct instead of making small talk with the interlocutor before bringing up the topic. They did not respond to the interlocutor's "How are you?" or "What's up," instead making a preemptive move to the reason for the call. He wrote that it might be harder for a native speaker with no experience talking with a non-native speaker in the task contexts to understand or accept such telephone opening styles. He mentioned that such trends or tendencies he observed in the two interactive tasks were somewhat the opposite of Korean culture, which is more indirect in terms of reporting some problems. Thus, he reasoned that sociolinguistic background knowledge was the most important part of knowledge for test-takers to perform well apart from language knowledge.

In terms of difficulty, he commented that eliciting more conversation from

the participants was the most difficult part to meet the desired 3-4 minutes. This might have resulted from the fact that the tasks were relatively easy or that test-takers were not accustomed to phone communication.

In regard to performance, he indicated that use of colloquial expressions and more fluent speech were the most noticeable features for higher levels. The test-takers at lower proficiency levels tended to use some awkward expressions and spoke with some long pauses. One common problem he noticed across their performances regardless of their levels was that they displayed some unnatural use of articles and some incorrect prepositions.

Finally, when asked to comment about the differences between the test-takers' performances and that of a native speaker, he mentioned that he would have expected a native speaker to explain or show that he or she attempted to keep the appointment. A friendly tone and doing a proper telephone opening sequence could be part of the criterion for this type of speaking assessment.

Chapter V

Discussion

This chapter discusses the major findings reported in the previous chapter in terms of 1) task-takers' performances in the composite scores, and 2) the language features. It also addresses test-takers' answers to the questionnaires and some of the raters' and the interlocutor's feedback to the short questionnaire with respect to the tasks and the rater reliability.

5.1 Task Effect in the Composite Scores

The higher correlation between the composite scores for the two different tests indicates that it is more likely that the test-takers who performed well on the non-interactive tests also performed well in the interactive test. Such results were the opposite of the findings by Helleck (2007), who reported that examinees received higher ratings on the monologic role-play task. Such high correlation of the performances on the conditions for the study implies that non-interactive role-play tasks could be used as measurement in that candidate's speaking ability is not that differently captured in the interactive role-play task. However, upon closer inspection, there were some participants whose performances were clearly different in terms of the composite scores. For instance, some participants outperformed in the non-interactive condition, while others scored higher in the interactive condition. Individual variations might have come into play for this.

5.2 Analysis of Language Samples

Higher mean scores found in the pronunciation/intonation and fluency criteria in the descriptive statistics suggests some differences in the language produced in the two conditions. Although the data were not analyzed with the aid of acoustic-phonetic measures, the length of time for each subject's performance in the two conditions was calculated, and the number of conversational turns for the interactive condition was also counted. Thus, the examination into the relationship between the length of time and the number of conversation turns provided some explanation about the differences in those criteria, although it is difficult to make a direct comparison, since the two task conditions do not share the same scoring rubrics due to the two task-dependent analytic criteria. This can be explained by the effects of planning. Planning might have affected the fluency in the non-interactive condition, where test-takers could benefit from preparing for their answers. This was supported by some participants who commented that the preparation time given for the interactive test was not helpful, since they could not use that time to anticipate the interlocutor's responses beforehand. However, these results were contrary to what was suggested by Ejzenberg (2000). She argued that L2 learners will receive higher scores in terms of fluency when they interact with a native speaker, compared to other conditions; fluency will be negatively affected in the non-interactive task condition due to the high cognitive demands on the part of the speaker. Numerous studies concerning task effects have found that fluency can be displayed variously depending on

the properties of the task. (e.g. Bygate, 1996; Ejzenberg, 1992; Foster & Skehan, 1996; 1999). As evidence, the test-takers for the current study stated that they had problems in performing the interactive test because they had to not only think about what to say, searching for the appropriate words, but also that they were concerned with their pronunciation and whether or not the native speaker could understand them. Quick reactions required in the phone communication context might have affected the scores in the pronunciation/intonation and fluency criteria for Korean learners of English who were not accustomed to speaking in English on the phone. This clearly indicates that the language features of pronunciation/intonation and fluency (i.e. pauses, hesitation, and repetition) are areas where one's language competence, including interactive competence, is revealed in the interactive context. Some features of pronunciation/intonation and fluency, which are difficult to capture in the non-interactive context, can be assessed more accurately in the interactive context where quick responses are required, which is essential in real-time speaking communication.

5.3 Test-takers' Perceptions

Test-takers' answers to the questionnaire provide valuable insights to the overall test and future studies. As described in the Previous Results section above, the test-takers' reported anxiety levels as well as self-assessment rate were not that different across the two test conditions.

In terms of task difficulty, participants agreed that the tests were not difficult to perform. This might be explained by the topics and the intended

audience used for the current study, which were easily relatable to their daily lives as university students. With regard to task interestingness and authenticity, participants were positive about both tests, although there were higher scores on the interactive test. Also, they responded positively to the question asking about the extent to which their ability to speak can be demonstrated in the two test conditions, although slightly higher rates were observed on the interactive test. Backman (1990) defined authenticity as "a function of the interaction between the test taker and the test task, (p.317)" and so, the negotiation of meaning is a crucial constituent in test authenticity. He also argued that test-takers' perceptions of authenticity of assessment tasks in a particular TLU domain would serve as a key role in their performance on those tasks because how they perceive authenticity will greatly influence their performance on a task. In this vein, it is encouraging and meaningful that the test-takers perceived the interactive phone communicative task as authentic and more representative of real-life situations, which is a sign that more authentic tasks need to be designed and used so that test-takers could be engaged in communicative language use.

It is noteworthy that the test-takers strongly disagreed that their performance would have been better if they had talked to a different interlocutor, in that this demonstrates that the interlocutor did a proper job of being a conversational partner as a hypothetical friend, and they did not experience unfairness when being tested in the interactive tasks. However, about half of the test-takers agreed with the statement that they would have performed better if they had talked to the interlocutor face-to-face. This is an

indication that communication was somewhat difficult in the phone communication context, where nonverbal communication cues such as facial expressions or gestures were not available as opposed to communication in person. This suggests that phone communication would be more challenging for learners, although it represents an essential skill needed in real-life situations. Thus, tasks to represent phone communication in a real-life domain would be helpful for both learners and test-users.

5.4 Raters' Feedback

The raters' feedback provided valuable insights with respect to the scoring rubric. The two raters' overall opinions and impressions on the participants' performances were positive, and their perceptions were revealed in the composite mean scores of the two tests. Both raters asserted that they had some difficulties with rating the pragmatic meanings and interactive communication criteria due to ambiguity in terms of discerning these two measures. They pointed out some problems in the analytic rating criteria regarding the control of pragmatic meanings and interactional communication in the interactive test due to the overlapping features in the two descriptors. This provides crucial feedback that should be taken into consideration for future studies. Such comments were supported by the correlation coefficients. Although in general, significantly high correlation coefficients were found in the rater reliability, a closer look at the analytic rating criteria reveals that the lowest correlation coefficient values were reported in the interactional communication criterion. Accordingly, the lowest Kappa coefficient values

were also found in the interactional communication criterion. All these reveal the need for more specificity of the descriptors in the rubric.

5.5 The Interlocutor's Feedback

Among the answers responded to in the short questionnaire by the interlocutor, one important observation he made is that the task-takers' telephone opening sequence was different from the norm in Western culture. Most of the participants moved directly to the reason for the call by omitting their relevant response to the interlocutor's "How are you?" or "What's up?".

Caution should be taken in interpreting such distinct patterns, since mobile telephone call openings can be different from landline telephone call openings. These days, the caller is identified on the mobile phone screen so that the recipient can know who is calling before answering the phone. (Arminonen & Leinonen, 2006). In addition, it could be that task effect in this study played a role in jumping into the topic without proper greetings or self-presentation. The participants were presented with what to accomplish—reporting a problem and suggesting solutions in a given role to resolve it—in the prompt, , and also knew who was going to answer their calls in the given context. Another possibility is that the test-takers were embarrassed about doing the task, acting like a friend with someone whom they do not know.

Nevertheless, the observed pattern of the participants not reciprocating with the interlocutor seemed to be problematic from a second-language pedagogical perspective. Pragmatic knowledge in relation to taking a phone call or making a phone call needs to be taught, particularly to lower-level

proficiency students. The interlocutor also pointed out that it might be difficult for those who have no experience in talking with non-native speakers of English to understand or accept such behavior if it was in a real-life context. These findings indicate there is a need for further investigation.

Chapter VI

Conclusion

6.1 Conclusions and Implications

The findings of this study demonstrated that the test-takers' performances were not different between the non-interactive and interactive conditions. However, the mean scores in both the pronunciation/intonation criterion and the fluency criterion were higher in the non-interactive test than in the interactive test, and the nature of fluency was reflected in the number of conversation turns in the interactive test. Also, the results of the study suggest that the test-takers preferred to do the interactive tasks. In this sense, the combination of the monologic and dialogic levels of tasks would more accurately evaluate one's speaking ability with respect to the pronunciation/intonation and fluency components, in particular.

The interactive tasks used for this study attempted to elicit dialogic speech samples in the semi-direct assessment condition. They seemed not only practical but also authentic, since real conversation ability was assessed through the mobile phone. Based on the findings of the study, it appears that pronunciation/intonation and fluency are the features that could be captured differently in the non-interactive and interactive conditions. One major implication is the need to develop and apply tasks that can elicit authentic

dialogic-level interaction in the semi-direct speaking assessment. At present, tasks to elicit monologic speech samples in the role-play context of leaving a voice mail message or talking to a hypothetical audience are widely used in computer-mediated speaking assessment (e.g. the TOEIC Speaking test and the ACTFL OPIc). The newly developed exam, the National English Ability Test (NEAT) for adults, also includes such task types (i.e. leaving a voice mail) in the speaking section. Thus, the addition of a task that could evaluate one's interactive communication competency could lead to positive washback in the Korean EFL education context because learners would be motivated to study and practice speaking skills.

With the application of mobile phones or VoIP, more authentic and more accurate assessment could be possible in the semi-direct speaking assessment. The current study is a prototype task study. More research is needed with regard to development of tasks to evaluate real ability to speak in order to shed light on speaking assessment. It would be of interest to investigate test-takers' performances and their perceptions in various TLU domains in a phone communication context. Further research is needed to devise effective tasks capable of enhancing validity, authenticity and reliability and thus provide a more accurate picture of learners' speaking ability.

6.2 Limitations and Future Studies

There were a number of limitations in terms of the methodology and analysis of data collected for the study. First of all, the study was not

administered in a fully semi-direct mode due to logistic limitations, which suggests that the dichotomy of direct and semi-direct speaking assessment should be viewed on a continuum rather than as two separate extremes. Secondly, the intended audience and the recipient in both task conditions were limited to a friend, which might have affected the degree of difficulty. This limits the analysis of data to the scores in the performance on this single relationship between friends. The tasks with a variety of TLU domains that elicit various functional languages such as complaining, persuading, and giving advice could provide a more accurate measure of test-takers' speaking performances in a phone communication context and of score reliability. Another limitation was that lower level test-takers' performances on the two tests were not taken into consideration. Given that the test-takers' performances did not differ across the two conditions in this study, the inclusion of that group would have yielded different results in terms of task difficulty.

The purpose of this study was to investigate how test-takers perform differently in the existing monologic tasks and the dialogic tasks where interaction is involved, and thus the specific tasks and scoring rubric to target those tasks used in this study had to be created newly by the researcher. Due to the nascent nature of such a scoring rubric, some limitations were found. Although the comparison was made based on the composite of the five analytic scores in each task, analytic criteria in the two conditions were not the same in terms of comparing test-takers' performances. This might have affected the results of the study. Furthermore, while high correlations between

the two raters were reported overall, rater score disagreement was found in some analytic criteria such as task achievement and discourse management in the non-interactive test, and interactional communication in the interactive test. Not all the speech samples were qualitatively analyzed in terms of discourse analysis. Qualitative analysis of the speech samples could yield results that may help to improve the scoring rubrics by identifying the elements in speech which were not captured by the scoring descriptors. Given that a scoring rubric is a fundamental area in fair assessment since it informs the construct of speaking ability being measured, qualitative analysis of speech data will contribute to more accurate measure of speaking performance and provision of diagnostic information.

References

Audacity (version 2.0.0). (2012). Retrieved March, 14, 2012 from http:audacity.sourceforge.net/

Arminonen, I., &, Leinonen, M. (2006). Mobile phone call openings: tailoring answers to personalized summonses. *Discourse studies*, *8*(3), 339-368.

Bachman, L.F. (1990). *Fundamental considerations in language testing.* Oxford University Press.

Bachman L.F. and Palmer A.S. (1996). *Language testing in practice.* Oxford University Press.

Brown, A., N. Iwashita, & T. McNamara. (2005). *An examination of rater orientations and test- taker performance on English for academic purposes Speaking Tasks*. *RESEARCH REPORT-EDUCATIONAL TESTING SERVICE PRINCETON RR, 5*.

Brown, A., & Lumley, T. (1997). Interviewer variability in specific-purpose language performance tests. *Current developments and alternatives in language assessment*, 137-150.

Brown, A. & Hill, K. (1998). Interviewer style and candidate performance in the IELTS oral interview. *IELTS research reports*, *1*, 1-19.

Bygate, M. (1996). Effects of task repetition: appraising learners' performances on tasks. *Challenge and change in language teaching*, 136-146

Canale, M., & Swain, M. (1980). Theoretical bases of communicative approaches to second language teaching and testing. *Applied Linguistics* 1, 1-47

Carroll, J. B. (1983). Psychometric theory and language testing. *Issues in language testing research*, 80-107.

Chapelle, C., & Douglas, D. (2006). *Assessing language through computer technology*. Cambridge University Press

Chun, C. (2006). An analysis of a Language Test for Employment: The authenticity of the PhonePass Test, *Language Assessment Quarterly*, 3(3), 295-306

Cohen, L & Manion, L. (1980). *Research Methods in Education*, (2nd ed.). Dover, NH: Croom Helm.

Cumming, A., Grant, L. , Mulcahy-Ernt, P. , & Powers D.(2004). Study of speaking and writing prototype tasks for a new TOEFL, Language Testing

Douglas, D., & Hegelheimer, V. (2007). Assessing language using computer technology. *Annual Review of Applied Linguistics*, *27*, 115.

Ejzenberg, R. (1992). *Understanding nonnative oral fluency: The role of task structure and discourse variability*. Ann Arbor, MI: University Microfilms International.

Ejzenberg, R. (2000). The juggling act of oral fluency: A psycho-social linguistic metaphor. In H. Riggenbach (Ed.) *Perspectives on Fluency*. (pp. 287-313) Ann Arbor: University of Michigan Press.

Foster, P., & Skehan, P. (1996). The influence of planning and task type on second language performance. Studies *in Second Language Acquisition*, 18(3), 299–323.

Foster, P., & Skehan, P. (1999). The influence of sources of planning and f

ocus of planning on task-based performance. *Language Teaching Research, 3*(3), 215–247.

Fulcher, G. (1997). The testing of speaking in a second language. *Encyclopedia of language and education*, *7*, 75-85.

Fulcher, G. (2003). *Testing Second Language Speaking*. Pearson Education.

Fulcher G. and Reiter, R. M. (2003).Task difficulty in speaking tests *Language Testing*, 20(3), 321-344.

Galaczi, E. D. (2010). Face-to-face and computer-based assessment of speaking: Challenges and opportunities. *Computer-based Assessment (CBA) of Foreign Language Speaking Skills*, 29.

Halleck, G. (2007). Symposium article: Data generation through role-play: Assessing oral proficiency, *Simulation & Gaming 38*(1), *91-106*

IBM Corp. (2011). IBM SPSS Statistics for Windows, Version 20.0. Armonk, NY: IBM Corp.

Iwashita, N., McNamara, T. & Elder, C. (2001). Can we predict task difficulty in an oral proficiency test? Exploring the potential of an information processing approach to task design. *Language Learning 51*(3), 401-36.

Jamieson, J. (2005). Trends in computer-based second language assessment. *Annual Review of Applied Linguistics*, *25*(1), 228-242.

Kim, H. J. (2011). Investigating effects of tasks on examinee performance in a computer-delivered speaking test. *Multimedia-assisted language learning*, *14*(1), 65-93.

Kormos, J.(1999). Simulating conversations in oral-proficiency assessment: a conversational analysis of role-plays and non-scripted inverviews in

language exams. *Language Testing 16*(2), 163-188.

Kuo, J.,and Jiang, X. (1997). Assessing the assessments: The OPI and the SOPI. *Foreign Language Annals, 30*(4), 503-512.

Landis, J. & Koch G. (1997). The measurement of observer agreement for categorical data. *Biometrics*, 159-174.

Lazaraton, A. (1992). The structural organization of a language interview: a conversation analytic perspective. *System 20*(3), 373-86

Lazaraton, A. (1996) Interlocutor support in oral proficiency interview: The case of CASE. *Language Testing, 13*(2), 151-172.

Lee.,Y.W. (2006). Dependability of scores for a new ESL speaking assessment, *Language Testing, 23*(2), 131-166.

Luoma, S. (2004). *Assessing Speaking*. Cambridge language assessment series.

Lumley, T. & O'Sullivan, B. (2005). The effect of test-taker gender, audience and topic on task performance in tape-mediate assessment of speaking. *Language Testing, 22*(4), 415-437.

Malabonga, V., Kenyon, D.M., & Carpenter, H. (2005). Self-assessment, preparation, and response time on a computerized oral proficiency test., *Language testing, 22*(1), 59-92.

McNamara, T. F. (1996). *Measuring Second Language Performance*. London: Longman.

McNamara, T. (1997). 'Interaction' in second language performance a ssessment: Whose performance? , 1. *Applied Linguistics, 18(4)*, 446-466.

McNamara, T., & Lumely, T. (1997). *The effect of interlocutor and*

assessment mode variables in overseas assessments of speaking skills in occupational settings. Language Testing, 14(2), 140-156.

Milanovic, M., & Saville, N. (Eds.). (1996). *Performance Testing, Cognition and Assessment: Selected Papers from the 15th Language Research Testing Colloquium, Cambridge and Arnhem* (Vol. 3). Cambridge University Press.

Morton, J., Wigglesworth, G., & Wiliams, D. (1997). Approaches to the evaluation of the interviewer performance in oral interaction tests. *Access: Issues in English Language Test Design and Delivery. Sidney: National Centre for English Language Teaching and Research*, 175-196.

M. Celce-Murcia, Z. Dörnyei, & S. Thurrell (1995). Communicative competence: a pedagogically motivated model with content specifications. *Issues in Applied Linguistics 6*(2), 5-35.

Nakatsuhara, F. (2007). Developing a rating scale to assess English speaking skills of Japanese upper-secondary students. *Essex Graduate Student Papers in Language and Linguistics, 9,* 83-103.

O'Sullivan, B., Weir, C., & Saville, N. (2002). Using observation checklists to validate speaking-test tasks. *Language Testing, 19*(1), 33-56.

Okada., Y. (2010). Role-play in oral proficiency interview: Interactive footing and interactional competencies. *Journal of Pragmatics, 42(6),* 1647-1668.

Purpura, J. E.(2004). *Assessing grammar*. Cambridge: Cambridge University Press

Qian, D. (2009). Comparing direct and semi-direct modes for speaking assessment: Affective effects on test takers. *Language Assessment Quarterly*, *6*(2), 113-125.

Reed, J. (1990). Providing relevant content in an EAP writing test. English for Specific Purposes 9(2), 109-121

Reed, J. & G. Halleck (1997). Probing above the ceiling in oral interviews. What's up there. *Current developments and alternatives in language assessment: Proceedings of LTRC* (Vol. 96, pp. 225-238).

Ross, S. & Berwick, R. (1992). The discourse of accommodation in oral Proficiency interviews. *Studies in Second Language Acquisition, 14*(2), 159-176.

Rosendale, D. (1989). Role-play as a data-generation method. *Simulation gaming, 20(4)*, 487-492.

Skehan, P. (1998a). A cognitive approach to language learning. Oxford; Oxford University Press.

Skehan, P. (1998b). Processing perspectives to second language development, instruction, performance and assessment. *Thames Valley Working Papers in Applied Linguistics* 4, 70-88.

Shohamy, E. (1983). The Stability of oral proficiency assessment on the oral interview testing procedures. *Language Learning 33*(4), 527-540

Shohamy, E., Reves, T. & Bejerano, Y. (1986). Introducing a new comprehensive test of oral proficienc. *English Language Teaching Journal* 40 (3), 212-220

Shohamy, E. (1988). A proposed framework for testing the oral language of

second/foreign language learners. *Studies in Second Language Acquisition*, *10*(02), 165-179.

Sim, J. & Wright, C. (2005). The kappa statistic in reliability studies: Use, interpretation, and sample size requirements. *The Journal of the American Physical Therapy Association, 85* (3), 257-268.

Spearman, C. (1910). The proof and measurement of the association between two things. *American Journal of Psychology, 15*(1), 72-101.

Stansfield, C.W. (1990). An evaluation of simulated oral proficiency interviews as measures of oral proficiency. *Georgetown Roundtable on Languages and Linguistics 1990*, 228-234.

Stansfield, C.W.,&Kenyon, D.M. (1988). *Development of the Portuguese Speaking Test*. Washington, DC: Centre for Applied Linguistics.

Stansfield, C. W., & Kenyon, D. M. (1992). Research on the comparability of the oral proficiency interview and the simulated oral proficiency interview.*System*, *20*(3), 347-364.

The Cambridge ESOL Bulats Online Speaking Tests, from http://www. bulats.org.

The ACTFL OPIc Speaking Test, from a validated computer-delivered oral proficiency assessment (2009).

The TOEIC Speaking Test Speaking and writing sample tests (2008), from http://www.ets.org

The Versant English Test, from Versant English test: test description and validation summary (2008). Pearson Education, Inc.

The TOEFL iBT Test , from http://www. ets.org.

TEPS. (2009). from http://www.teps.or.kr/

van Lier, L. (1989). Reeling, writhing, drawling, stretching, and fainting in coils: oral proficiency interviews as conversations. *TESOL Quarterly*, 23 (3), 489-508.

Wong, J., and Waring, H.Z. (2010). *Conversation Analysis and Second Language Pedagogy: A guide for ESL/EFL teachers.* Routledge.

Weir, C.J. (1993). Understanding and developing language tests. Prentice Hall.

Wesche, B. (1987). Second language performance testing: the Ontario test of ESL as an example. *Language Testing*, *4*(1), 28-47

Wigglesworth, G. (1997). An investigation of planning time and proficiency level on oral test discourse. *Language Testing*, *14*(1), 85-106.

Young, R. (1999). Sociolinguistic approaches to SLA. *Annual review of applied linguistics*, *19*, 105-134.

Young, R. & Milanovic, M. (1992). Discourse variation in oral proficiency interviews. *Studies in Second Language Acquisition*, *14*, 403-424

Young, R., & He, A. W. (Eds.). (1998). *Talking and testing: Discourse approaches to the assessment of oral proficiency* (Vol. 14). John Benjamins Publishing.

Appendix A

Individual Test

In this part of the test, you will be given two situations where you will have to leave a message on the phone. Imagine you are in each situation and leave appropriate messages. You will have 30 seconds to prepare and 90 seconds to respond to each.

Task 1,

Situation:

Imagine you have a foreign friend named Alex, who is going to visit you in Korea next week. You promised to be a tour guide for your friend. However, you have just found out that you must attend a big job fair and cannot be with your friend for one day. Your friend is not answering the phone, so you must leave a message to explain the problem and suggest solutions.

Task:

- explain the problem in detail

- make two or three suggestions that can help resolve the problem

Task 2

Situation:

Imagine you bought two tickets to a musical for yourself and your best friend. The musical is scheduled to be shown at 8:00 tomorrow night. However, you cannot go because you have to go on a business trip tomorrow morning. Your friend is not answering the phone, so you must leave a message to explain the problem to your friend.

Task:

- explain the problem in detail

- make two or three suggestions that can help resolve the problem.

Interactive Test

In this part of the test, you will be given two situations where you will interact with a native speaker of English on the phone. In the first situation, you will be asked to report a problem and make two or three suggestions to solve the problem. You need to discuss the problem with the speaker at the other end of the line in order to solve it. Imagine you are in this situation and act it out. You will be given 30 seconds to prepare, and have a maximum of four minutes to complete each task.

Task 1

Situation:

Imagine you have planned a trip with your friend for this coming weekend. However, you just found out that you have to attend an important meeting for a team project. The grade on the project will affect your final report. You are about to call your friend to explain the problem and discuss rescheduling your trip together. Refer to the schedule below to help you complete the task.

Task:

- explain the problem
- apologize
- suggest two or three options
- reach an agreement

<Native speaker's prompt>

Situation:

Imagine you have planned a trip with your friend for this coming weekend. However, your friend cannot go on the trip because something urgent has come up. You are about to receive a call from your friend. Listen to your friend and discuss rescheduling the trip together. Refer to the schedule below to help you complete the task.

What you need to do:

- recognize the problem
- ask questions with regard to the problem
- reach an agreement

My schedule

March 2013

Sun	Mon	Tue	Wed	Thu	Fri	Sat
			1	2	3	4
5	6	7	8	9	10 Jeiu Island	11 Team project meeting
12	13	14 Part-time Job	15	16 Part-time Job	17 The musical "Wicked" 8:00 PM	18
19	20	21 Part-time Job	22	23 Part-time Job	24	25 Cousin's wedding 12:00 PM
26	27	28 Part-time Job	29	30 Part-time Job	31	

NS's schedule

March 2013						
Sun	**Mon**	**Tue**	**Wed**	**Thu**	**Fri**	**Sat**
			1	2	3	4
5	6	7	8	9	10 Jeju Island	11
12	13	14	15 Volunteer work	16	17	18 Visiting Grand-parents
19	20	21 Seminar 9:15 AM-4:00 PM	22	23	24	25
26	27	28 Study group 9:30AM (3 hours)	29	30 Study group 9:30AM (3 hours)	31	

Task 2

Situation:

You know that your roommate is planning a birthday party for you with your friends on this coming Saturday. However, you just found out that you will be out of town with your parents on Saturday because they are visiting you. You have to call your roommate to discuss the birthday plan.

Task:

-explain the problem

-apologize

-give two or three alternatives

-reach an agreement

<Native speaker's prompt>

Situation:

Imagine that you have planned your roommate's birthday party and invited your friends to attend on this coming Saturday. You are about to receive a call from your roommate, who will report that something has come up on that day. Listen to your friend and discuss the problem.

What you need to do:

- recognize the problem

- ask some questions with regard to the problem

- reach an agreement

Appendix C

Questionnaire type A

Questionnaire for Group A

Please complete the following by inserting you answer or checking the appropriate box or the number on the scale that most accurately reflects your response to each question below.
다음에 나오는 각각의 질문에 답하시거나, 의견에 부합하는 숫자나 상자에 체크해 주십시오.

Background Information

Name *
이름이 무엇입니까?

Gender *
성별이 어떻게 됩니까?
- Male 남자
- Female 여자

Age *
나이가 어떻게 됩니까?

Major *
전공이 무엇입니까?

Have you ever lived in an English speaking country? *
영어 사용권 나라에서 거주한 적이 있습니까?
- None
- Less than 6 months
- More than 6 months but less than a year
- More than a year but less than 2 years
- More than 2 years

Page 2 — After page 1 Continue to next page

The voice mail task

Please complete the following by inserting your answer or checking the appropriate box or the number on the scale that most accurately reflects your response to each question below.
다음에 나오는 각각의 질문에 답하시거나, 의견에 부합하는 숫자나 상자에 체크해주십시오.

1. I felt nervous while I was doing the task. *
나는 이 시험을 보는 동안 긴장이 되었다.

	1	2	3	4	5	
Strongly diagree	○	○	○	○	○	Strongly agree

2. I believe I did well on the task. *
이 시험을 잘 본 것 같다.

	1	2	3	4	5	
Strongly disagree	○	○	○	○	○	Strongly agree

3. I felt the time allowed for the task was too short. *
시험 시간이 너무 짧다고 느꼈다.

	1	2	3	4	5	
Strongly disagree	○	○	○	○	○	Strongly agree

4. I felt I needed more preparation time. *
준비시간이 더 필요하다고 느꼈다.

	1	2	3	4	5	
Strongly disagree	○	○	○	○	○	Strongly agree

5. I understood what I was supposed to do in the task. *
이 시험에서 요구하는 바를 잘 이해했다.

	1	2	3	4	5	
Strongly disagree	○	○	○	○	○	Strongly agree

6. I felt that the task was too difficult. *
시험이 너무 어려웠다.

	1	2	3	4	5	
Strongly disagree	○	○	○	○	○	Strongly agree

7. I thought the task was interesting *
시험이 재미있었다.

	1	2	3	4	5	
Strongly disagree	○	○	○	○	○	Strongly agree

8. I think the task is more realistic and authentic *
시험이 좀 더 사실적 (실제 일어날 일에 가까운)인 것 같다.

	1	2	3	4	5	
Strongly disagree	○	○	○	○	○	Strongly agree

9. I believe that I was able to show my ability to speak English through the task. *
이 시험을 통해 나의 영어 말하기 능력을 보여줄 수 있었다고 생각한다.

	1	2	3	4	5	
Strongly disagree	○	○	○	○	○	Strongly agree

10. What do you think is the most difficult aspect of this type of task? Please explain why. *
이런 유형의 시험에서 가장 어려운 부분은 무엇이라 생각하는지 써 주세요.

Page 3 After page 2 Continue to next page

The phone task

Please complete the following by inserting your answer or checking the appropriate box or the number on the scale that most accurately reflects your response to each question below.
다음에 나오는 각각의 질문에 답하시거나, 의견에 부합하는 숫자나 상자에 체크해주십시오.

1. I felt nervous while I was doing the task. *
나는 이 시험을 보는 동안 긴장이 되었다.

	1	2	3	4	5	
Strongly disagree	○	○	○	○	○	Strongly agree

2. I believe I did well on the task. *
이 시험을 잘 본 것 같다.

	1	2	3	4	5	
Strongly disagree	○	○	○	○	○	Strongly agree

3. I felt the time allowed for the task was too short. *
시험 시간이 너무 짧다고 느꼈다.

	1	2	3	4	5	
Strongly disagree	○	○	○	○	○	Strongly agree

4. I felt I needed more preparation time. *
준비시간이 더 필요하다고 느꼈다.

	1	2	3	4	5	
Strongly disagree	○	○	○	○	○	Strongly agree

5. I understood what I was supposed to do in the task. *
시험에서 요구하는 바를 잘 이해했다.

	1	2	3	4	5	
Strongly disagree	○	○	○	○	○	Strongly agree

6. I felt the task was too difficult *
시험이 너무 어려웠다.

	1	2	3	4	5	
Strongly disagree	○	○	○	○	○	Strongly agree

I understood clearly what the person at the other end of the line was saying. *
7. 나는 상대방이 말하는 것을 잘 이해했다.

	1	2	3	4	5	
Strongly disagree	○	○	○	○	○	Strongly agree

8. I would have done better if I had talked with a different person. *
다른 사람과 대화했다면 더 잘 했을 것 같다.

	1	2	3	4	5	
Strongly disagree	○	○	○	○	○	Strongly agree

9. I would have done better if I had talked with the person face to face. *
사람을 직접 보고 얘기했다면, 더 잘했을 것 같다.

	1	2	3	4	5	
Strongly disagree	○	○	○	○	○	Strongly agree

10. I thought that the task was interesting. *
시험이 재미있었다.

	1	2	3	4	5	
Strongly disagree	○	○	○	○	○	Strongly agree

11. I think the task is more realistic and authentic. *
시험이 좀 더 사실적(실제 일어날 일에 가까운)인 것 같다.

	1	2	3	4	5	
Strongly disagree	○	○	○	○	○	Strongly agree

12. I believe I was able to show my ability to speak English through the phone task. *
이 시험을 통해 나의 영어 말하기 실력을 보여 줄 수 있다고 생각한다.

	1	2	3	4	5	
Strongly disagree	○	○	○	○	○	Strongly agree

13. What do you think is the most difficult aspect of this type of task? Please explain why. *
이런 유형의 시험의 가장 어려운 점은 무엇이라 생각하나요? 설명해 주세요.

ge 4 After page 3 Continue to

The voice mail task vs. The phone task

Please complete the following by checking the appropriate box or inserting your answer.
다음에 나오는 질문의 의견에 답하시거나 의견에 부합하는 상자에 체크해 주십시오.

1. Which task did you feel was more difficult to perform? *
이런 유형의 시험 중 어느 것이 더 어려웠나요?

- ☐ Leaving a voice mail message
- ☐ Talking on the phone

2. Which task did you feel more nervous taking? *
두 유형의 시험 중 어느 것을 볼 때 더 긴장되었나요?

- ☐ Leaving a voice mail message
- ☐ Talking on the phone

3. Which task do you feel would show your current ability to speak English in real life situations more accurately? *
두 유형의 시험 중 어느 것이 당신의 실제 상황에서 영어를 말할 수 있는 능력을 더 정확하게 보여 줄 수 있다고 생각하나요?

- ☐ Leaving a voice mail message
- ☐ Talking on the phone

4. Which task would you prefer to do on a test? Please explain why. *
어느 유형의 시험을 보는 것을 더 선호하십니까? 설명해 주세요.

Appendix D

Questionnaire type B

Questionnaire for Group B

Please complete the following by inserting your answer or checking the appropriate box or the number on the scale that most accurately reflects your response to each question below.
다음에 나오는 각각의 질문에 답하시거나, 의견에 부합하는 숫자나 상자에 체크해주십시오.

Background Information

Name *
이름이 무엇입니까?

Gender *
성별이 어떻게 됩니까?

- Male 남자
- Female 여자

Age *
나이가 어떻게 됩니까?

Major *
전공이 무엇입니까?

Have you ever lived in an English speaking country? *
영어 사용권 나라에서 거주한 적이 있습니까?

- None
- Less than 6 months
- More than 6 months but less than a year
- More than a year but less than 2 years
- More than 2 years

ıge 2 After page 1 **Continue to next page**

The phone task

Please complete the following by checking the number on the scale that most accurately reflects your response to each question or inserting your answer below.
다음 아래에 나오는 각각의 질문에 부합하는 숫자를 체크하거나 답해주십시오.

1. I felt nervous while I was doing the phone task. *
나는 이 시험을 보는 동안 긴장이 되었다.

	1	2	3	4	5	
Strongly disagree	○	○	○	○	○	Strongly agree

2. I believe I did well on the task. *
이 시험을 잘 본 것 같다.

	1	2	3	4	5	
Strongly disagree	○	○	○	○	○	Strongly agree

3. I felt the time allowed for the task was too short. *
시험 시간이 너무 짧게 느껴졌다.

	1	2	3	4	5	
Strongly disagree	○	○	○	○	○	Strongly agree

4. I felt I needed more preparation time. *
준비 시간이 더 필요하다고 느꼈다.

	1	2	3	4	5	
Strongly disagree	○	○	○	○	○	Strongly agree

5. I understood what I was supposed to do in the task. *
이 시험에서 요구하는 바를 잘 이해했다.

	1	2	3	4	5	
Strongly disagree	○	○	○	○	○	Strongly agree

6. I felt the task was too difficult. *
시험이 너무 어려웠다.

	1	2	3	4	5	
Strongly disagree	○	○	○	○	○	Strongly agree

7. I understood clearly what the person at the other end of the line was saying. *
나는 상대방이 말하는 것을 잘 이해했다.

	1	2	3	4	5	
Strongly disagree	○	○	○	○	○	Strongly agree

8. I would have done better if I had talked with a different person. *
다른 사람과 대화했다면 더 잘 했을 것 같다.

	1	2	3	4	5	
Strongly disagree	○	○	○	○	○	Strongly agree

9. I would have done better if I had talked with the person face to face. *
사람을 직접 보고 얘기했다면 더 잘 했을 것 같다.

	1	2	3	4	5	
Strongly disagree	○	○	○	○	○	Strongly agree

10. I thought that the task was interesting. *
시험이 재미있었다.

	1	2	3	4	5

Strongly disagree ○ ○ ○ ○ ○ Strongly agree

11. I think the task is more realistic and authentic. *
이 시험은 좀 더 사실적 (실제 일어날 일에 가까운)인 것 같다.

1 2 3 4 5

Strongly disagree ○ ○ ○ ○ ○ Strongly agree

12. I believe I was able to show my ability to speak English through the task. *
이 시험을 통해 나의 영어 말하기 실력을 보여 줄 수 있다고 생각한다.

1 2 3 4 5

Strongly disagree ○ ○ ○ ○ ○ Strongly agree

13. What do you think is the most difficult aspect of this type of task? Please explain why. *
이런 유형의 시험 중 가장 어려운 점은 무엇이라 생각하나요? 설명해 주세요.

ge 3 After page 2 Continue t

The voice mail task

Please complete the following by checking the appropriate number on the scale that most accurately reflects your response to each question or inserting your answer below.
다음 아래에 나오는 각각의 의견에 부합하는 숫자에 체크하시거나 질문에 답변해 주세요.

1. I felt nervous while I was doing the task. *
나는 이 시험을 보는 동안 긴장이 되었다.

1 2 3 4 5

Strongly disagree ○ ○ ○ ○ ○ Strongly agree

2. I believe I did well on the task. *
이 시험을 잘 본 것 같다.

1 2 3 4 5

Strongly disagree ○ ○ ○ ○ ○ Strongly agree

3. I felt the time allowed for the task was too short. *
시험시간이 너무 짧다고 느꼈다.

1 2 3 4 5

Strongly disagree ○ ○ ○ ○ ○ Strongly agree

4. I felt I needed more preparation time. *
준비 시간이 더 필요하다고 느꼈다.

1 2 3 4 5

Strongly disagree ○ ○ ○ ○ ○ Strongly agree

5. I understood what I was supposed to do in the task. *
이 시험에서 요구하는 바를 잘 이해했다.

	1	2	3	4	5	
Strongly disagree	○	○	○	○	○	Strongly agree

6. I thought that the task was too difficult. *
시험이 너무 어려웠다.

	1	2	3	4	5	
Strongly disagree	○	○	○	○	○	Strongly agree

7. I thought that the task was interesting. *
시험이 재미있었다.

	1	2	3	4	5	
Strongly disagree	○	○	○	○	○	Strongly agree

8. I think the task is more realistic and authentic. *
이 시험은 좀 더 사실적 (실제 일어날 일에 가까운)인 것 같다.

	1	2	3	4	5	
Strongly disagree	○	○	○	○	○	Strongly agree

9. I believe that I was able to show my ability to speak English through the voice mail task. *
이 시험을 통해 나의 영어 말하기 능력을 보여줄 수 있었다고 생각한다.

	1	2	3	4	5	
Strongly disagree	○	○	○	○	○	Strongly agree

10. What do you think is the most difficult aspect of this type of task? Please explain why. *
이런 유형의 시험에서 가장 어려운 부분은 무엇이라 생각하는지 써 주세요.

e 4 After page 3 Continue t

The phone task vs. The voice mail task

Please complete the following by checking the appropriate box or inserting your answer below.
다음 아래에 나오는 질문에 대해 자신의 의견에 부합하는 상자에 체크하거나 답변해 주세요.

1. Which test did you feel was more difficult to perform? *
두 유형의 시험 중 어느 것이 더 어렵게 느껴졌나요?

- ☐ Talking on the phone
- ☐ Leaving a voice mail message

2. Which task did you feel more nervous taking? *

두 유형의 시험 중 어느 것이 더 긴장이 되었나요?

- [ ] Talking on the phone
- [ ] Leaving a voice mail message

3. Which task do you feel would show your current ability to speak English in real life situations more accurately? *

두 유형의 시험 중 어느 것이 당신의 실제 상황에서 영어를 말할 수 있는 능력을 더 정확하게 보여 줄 수 있다고 생각하나요?

- [ ] Talking on the phone
- [ ] Leaving a voice mail message

4. Which task would you prefer to do on a test? Please explain why. *

어느 유형의 시험을 보는 것을 더 선호하십니까? 설명해 주세요.

Appendix E

Questionnaire for Rater

Questionnaire for raters

Prior to rating questions

1. What features or criteria do you think are the most important when rating speaking performance?

2. What features of candidates' language production do you think you are influenced by most when assessing speaking performance? (Are there any "favorite features" you think important, and thus are influenced by as compared with other features?)

Post-rating questions

3. What was your overall opinion about the tasks?

4. What was the most difficult aspect of scoring the speech samples for each task type (non-interactive vs. interactive)?

5. Did you find anything that was not captured in the scoring rubrics-which, as a result, made it difficult for you to rate the speech samples?

6. Did you find any interesting things or distinctive patterns in candidates' performance in each task type (non-interactive vs. interactive)?

7. Do you have any other comments that are not addressed in the previous questions?

Appendix F

Interlocutor's feedback

1. What were the difficult parts in participating as the interlocutor in this role-play type of phone communication task?

2. Do you think that the difficult things have been resulted from the task itself or test-takers' proficiency /knowledge/ability to speak to perform this type of task?

3. You have talked with the subjects whose TEPS scores were ranged from 700 to 930. What were the noticeable features/criteria between high proficiency test-takers and intermediate proficiency test-takers?

4. Have you found some patterns shown in Korean test-takers' performances regardless of their proficiency?

5. What do you think were the distinctive differences between Korean test-takers' performances and expected native speakers' performances in this context of reporting a problem and suggesting some suggestions to a friend by phone?

6. Most test-takers did not properly do the telephone opening sequence that is the norm to your culture, jumping to reporting the problem, not saying hi/how are you or introducing themselves. Do you think "friendliness" be considered as important as sub- criterion under pragmatics/sociolinguistics in this context?

Appendix G

Analytic Scale for the non-interactive Test

	5	4	3	2	1
Task achievement And discourse management	The speaker provides appropriate rationale for solutions with a clear understanding of the situation, and successfully addresses the task, completing all the points discussed in the prompt with details and full elaboration required by the task - Completely coherent	The speaker's performance contains some elements from the 5 point descriptor and some elements from the 3 point descriptor. -fairly well elaborated as required by the task --Generally coherent	The speaker provides some form of rationale for solutions, but some minor misunderstanding of points OR lack of details OR leaving out of some points is displayed. -at times unclear -inadequately elaborated and/or at times irrelevant elaboration to the task - At times incoherent	The speaker's performance contains some elements from the 3 point descriptor and some elements from the 1 point descriptor. -generally unclear -poorly elaborated --Often incoherent	The speaker fails to provide rationale for solutions. Most of the response is unrelated to the task due to major incomprehension/ misunderstanding of the situation and the prompt. -generally unclear -poorly elaborated --Generally incoherent
Pronunciation and Intonation	The response is very intelligible and sustained. Articulation of	Speaks with some noticeable L1 prosodic features and individual sounds. However, the	Speaks with marked L1 prosodic features and individual sounds. The response occasionally	Speaks with strong L1-like pronunciation and intonation. The response	Speaks with very strong L1-like pronunciation and intonation. The

	individual sounds is very clear, although occasionally some L 1 - influenced sounds are still present; speaks mostly with appropriate word-stress/rhythm and adequate intonation.	response is generally intelligible and sustained, easy to understand, with appropriate rhythm and intonation- that does not require much listener effort.	requires listener effort to understand the speech, but does not interfere with comprehending.	frequently requires listener effort to understand the speech, and occasionally interferes with comprehending.	response requires much listener effort to understand speech, and almost always interferes with comprehending
Fluency	The speaker has almost natural speed with a well-paced flow. There may be some natural pauses when looking for language.	Pace may vary at times, being slow with a few occasional pauses when looking for appropriate language. However, it does not require unreasonable patience on the part of the listener.	Speech is slow, with several extended or unnatural pauses when looking for appropriate language. It requires occasional unreasonable patience on the part of the listener, but does not interfere with comprehending	Speech is very slow and choppy at some times. It has frequent extended or unnatural pauses, and repetitions of language. It frequently requires unreasonable patience on the part of the listener, and occasionally interferes with comprehending	Speech is very slow and choppy at almost all times. Socially inappropriate lengthy pauses and frequent repetitions make speech almost impossible to follow, interfering with comprehending.
Control of grammatical	The speaker uses a wide range of structures (this includes use of subordinate clauses and relative clauses, etc.) with few	The speaker uses a relatively wide range of structures, though there are some noticeable errors, in particular when using more complex	The speaker uses a narrow range of structures, which are mostly basic and simple sentences. When attempting to use complex	The speaker uses a limited range of structures and makes frequent errors, that occasionally interfere with comprehending. The speaker frequently	The speaker uses a very limited variety of structures, speaking in few full sentences, but mainly with isolated words and phrases. Almost entirely

forms **:** **-use accurate, diverse and complex grammatical forms** **- accurate use of morphological irregularity, formulaic forms, prepositions, mood, voice, modality, logical connectors, cohesive devices**	noticeable errors. The speaker use appropriate and sufficient range of vocabulary to deal with the situation	structures, though this does not interfere with comprehending. The speaker generally uses an appropriate range of vocabulary to deal with the situation, although some inaccurate usages are shown.	structures, several marked inaccuracies and frequent minor errors are present. The speaker occasionally uses inaccurate vocabulary to deal with the situation, and a limited range of vocabulary may interfere with comprehending	uses inaccurate vocabulary or awkward expressions to deal with the situation. The speaker uses a limited range of vocabulary occasionally interferes with comprehending.	inaccurate grammar interferes with comprehending at nearly all times. The speaker uses a very limited range of vocabulary often interferes with comprehending.
Control of pragmatic meanings: **-Sociolinguistic appropriateness** **--Socio-cultural appropriateness**	Completely appropriate openings and closings i.e. friendly tone to a friend i.e show apologetic attitude	Generally appropriate openings and closings	At times inappropriate openings and closings At times inappropriate	Often inappropriate openings and closings	Generally inappropriate openings and closings

Appendix H

Analytic Scale for the Interactive Test

	5	4	3	2	1
Pronunciation and Intonation	The response is very intelligible and sustained. Articulation of individual sounds is very clear, although occasionally some L 1 - influenced sounds are still present; speaks mostly with appropriate word-stress/rhythm and adequate intonation.	Speaks with some noticeable L1 prosodic features and individual sounds. However, the response is generally intelligible and sustained, easy to understand, with appropriate rhythm and intonation- that does not require much listener effort.	Speaks with marked L1 prosodic features and individual sounds. The response occasionally requires listener effort to understand the speech, but does not interfere with comprehending.	Speaks with strong L1-like pronunciation and intonation. The response frequently requires listener effort to understand the speech, and occasionally interferes with comprehending.	Speaks with very strong L1-like pronunciation and intonation. The response requires much listener effort to understand speech, and almost always interferes with comprehending
Fluency	The speaker has almost natural speed with a well-paced flow. There may be some natural pauses when looking for language.	Pace may vary at times, being slow with a few occasional pauses when looking for appropriate language. However, it does not require unreasonable patience on the part of	Speech is slow, with several extended or unnatural pauses when looking for appropriate language. It requires occasional unreasonable patience on the part of the	Speech is very slow and choppy at some times. It has frequent extended or unnatural pauses, and repetitions of language. It frequently requires unreasonable patience	Speech is very slow and choppy at almost all times. Socially inappropriate lengthy pauses and frequent repetitions make speech almost impossible to follow,

		the listener.	listener, but does not interfere with comprehending	on the part of the listener, and occasionally interferes with comprehending	interfering with comprehending.
Control of grammatical forms : **-use accurate, diverse and complex grammatical forms** **- accurate use of morphological irregularity, formulaic forms, prepositions, mood, voice, modality, logical connectors, cohesive devices**	The speaker uses a wide range of structures (this includes use of subordinate clauses and relative clauses, etc.) with few noticeable errors. The speaker use appropriate and sufficient range of vocabulary to deal with the situation	The speaker uses a relatively wide range of structures, though there are some noticeable errors, in particular when using more complex structures, though this does not interfere with comprehending. The speaker generally uses an appropriate range of vocabulary to deal with the situation, although some inaccurate usages are shown.	The speaker uses a narrow range of structures, which are mostly basic and simple sentences. When attempting to use complex structures, several marked inaccuracies and frequent minor errors are present. The speaker occasionally uses inaccurate vocabulary to deal with the situation, and a limited range of vocabulary may interfere with comprehending	The speaker uses a limited range of structures and makes frequent errors, that occasionally interfere with comprehending. The speaker frequently uses inaccurate vocabulary or awkward expressions to deal with the situation. The speaker uses a limited range of vocabulary occasionally interferes with comprehending.	The speaker uses a very limited variety of structures, speaking in few full sentences, but mainly with isolated words and phrases. Almost entirely inaccurate grammar interferes with comprehending at nearly all times. The speaker uses a very limited range of vocabulary often interferes with comprehending.
Control of pragmatic meanings: **-Sociolinguistic appropriateness** **-Socio-cultural**	Completely appropriate : distinguish register-friend well -Phone opening	Generally appropriate	At times inappropriate At times inappropriate	Often inappropriate.	Generally inappropriate

appropriateness	sequence/ closing sequence; not jump into the conversation abruptly, then move to topic development -show apologetic attitude - Although the first attempt to solve the problem failed, test-takers try to negotiate with the speaker actively, -Initiate and lead the conversation, not making the speaker dominate it (not requiring more talk on the part of NS)				
Interactional Communication	The speaker is almost entirely effective at communicating with the partner, both actively and receptively. Fully engaged in turn-taking and contributes to interactive	The speaker generally communicates effectively with the partner, through appropriately participating in turn-taking; attempts to contribute to interactive	Communication is sometimes effective, but occasionally inefficient because of some difficulty in understanding the partner and some unsuccessful attempts to negotiate meanings	Communication is neither effective nor meaningful because of difficulty in understanding the partner. Interacts mostly with monologue and simple responses, but is not	Communication is ineffective and inefficient because of only simple responses and a lack of ability to maintain or develop the interaction

	communication by negotiating meanings appropriately for the situation (e.g. initiates the conversation, asks questions, responds to questions, asks for clarification, responds to requests for clarification, terminates the conversation, etc.).	communication by negotiating meanings in some -but not all the occasions		fully engaged in negotiating meanings. There may be some attempts to negotiate meanings, but these are mostly unsuccessful.	

Appendix I

Consent Form

Thesis Study Consent Form

Title of the study: Examining Test-takers' Performance on Telephone Communication Task

I understand that the purpose of this study is to examine test-takers' performances on two types of tasks and test-takers' perception to each task. My participation in this study will entail performing four speaking task and filling out the questionnaire with regard the testing experience and tasks. The study involves two sessions testing comprised of two tasks in each testing; in the first session, I will be asked to complete a background questionnaire, perform two tasks, and read and rate the questionnaire with regard to experiences in the test. In the second session, I will be asked to perform two tasks, and read and rate the questionnaire with regard to experiences in the test. I understand that my responses are recorded, and that my oral and written data will be used for linguistic analysis. My responses will not be released to any person or institution apart from the person who is conducting the study. I understand that I will not be identified by name in any report of the study results. For participating in this study, I will be paid 10,000 won for completing the two sessions.

Signature:

_____

Date:

_____

Name

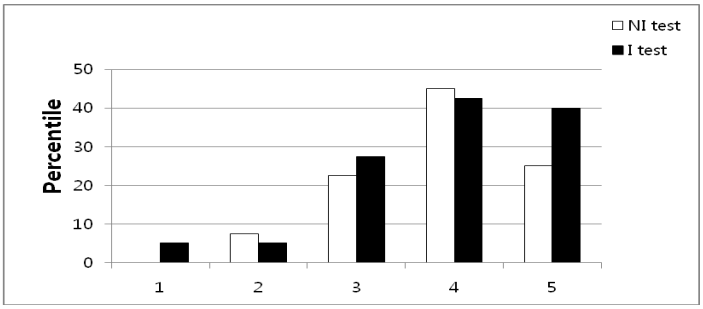
_____

Appendix I

Questions for the non-interactive and interactive tasks

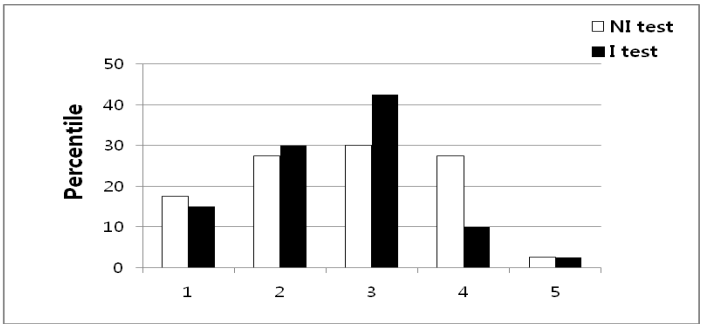
Question2.

I believe I did well on the task.



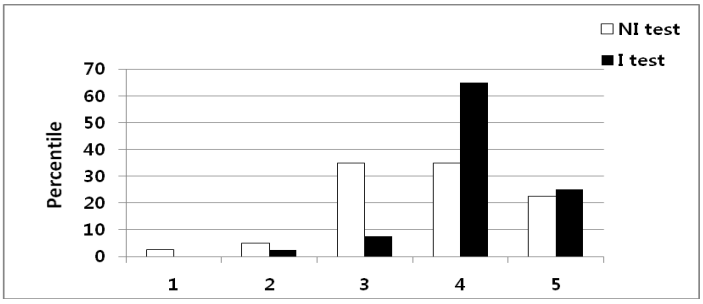
Question6

I felt that the task was too difficult.



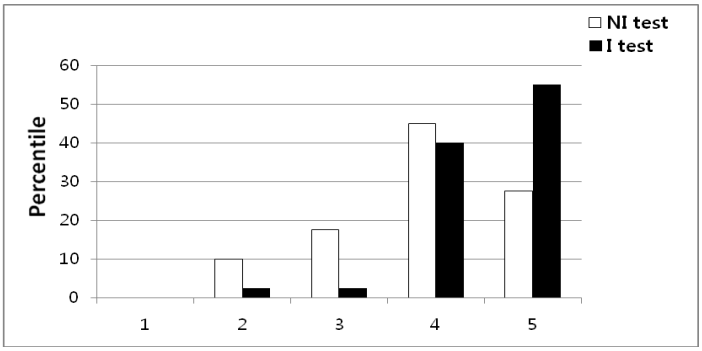
Question 7

I thought that the task was interesting.



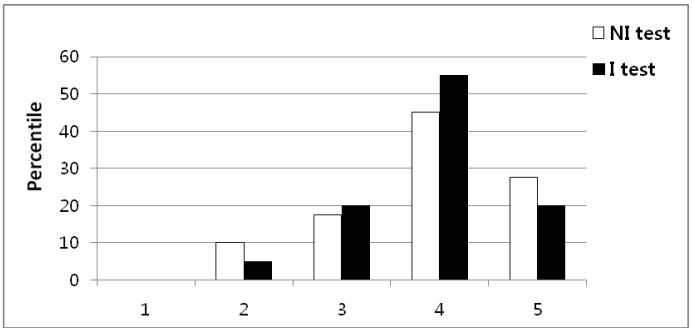
Question8

I think that the task is realistic and authentic.



Question 9

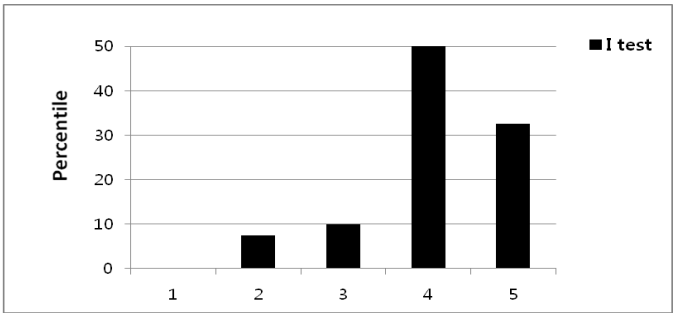
I believe that I was able to show my ability to speak English through the task.



<u>Questions for the interactive test only</u>

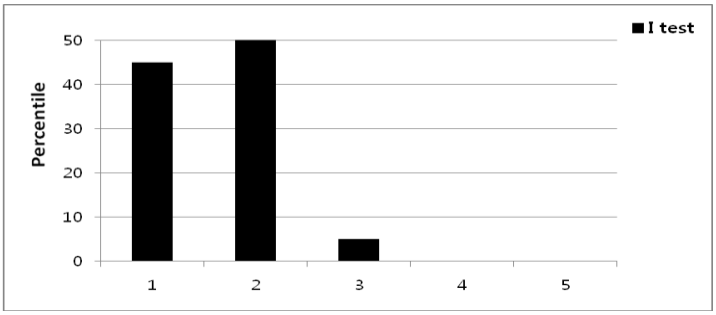
Question1

I understood clearly what the person at the other end of the line was saying.



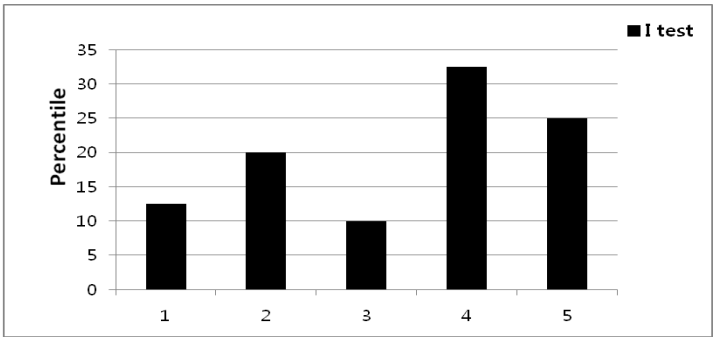
Question 2

I would have done better if I had talked with a different person.



Question 3

I would have done better if I had talked with the person face to face.



국문초록

EFL 영어말하기 평가 상황에서 상호작용적 전화 대화과제의 활용가능성 연구

정루미
서울대학교 대학원
영어영문학과 영어학 전공

기술매개 말하기평가는 시험 전달의 용이함과 높은 성적 신뢰도 때문에 최근에 널리 이용되는 추세이다. 하지만, 부정적인 측면에서 볼 때, 준직접적 말하기 시험은 상호작용의 부재로 인해 좁은 언어 기술만 평가될 수 있는 점을 비판 받아왔다. 기술 관련 직접 말하기 시험에서도 휴대전화나 음성인터넷 프로토콜을 활용함으로써 상호작용 영역을 평가할 수 있을 거라 예상되지만, 준직접적 말하기 시험에 상호작용적 특성을 시행하는 것에 대한 연구는 거의 전무하다. 따라서, 본 연구는 EFL 환경의 영어평가상황에서 상호작용이 가능한 휴대전화대화 과제를 활용하는 것에 대한 가능성을 연구했다.

44명의 영어를 외국어로 배운 한국인 대학생들이 연구자가 만든 말하기 과제를 수행하도록 모집되었다. 말하기 수행능력을 평가하기 위해서, 롤플레이 형식의 실제 상대자와 대화를 하는 휴대전화 대화 과제와 음성메세지를 남겨야하는 혼자 말하는 상황의 과제가 이용되었다. 두 명의 채점자가 다섯 개의 영역의 분석적 기준에 근거하여 시험자의 말하기 수행을 평가했다. 또한, 시험자와 채점자가 어떻게 말하기 과제를 인식하는지를 알아보기 위하여 설문조사도 실시하였다.

연구결과는 두 유형의 과제에서 높은 수준의 채점자간 신뢰도가 발견되었지만, 분석적인 영역을 살펴보면, 혼자 말하기 과제에서는 과제수행 및 담화 관리영역, 대화 과제에서는 상호적인 의사소통영역에서 다소 불일치가 발견되었다. 항목간과 기준 시험 성적간 발견된 강한 양의 상관관계는 타당도를 주장하는 근거이다. 두 유형의 과제에서 나타나는 주목할만한 특징을 분석하기 위해, 각 시험자의 모든 과제에서 말했던 시간과, 대화과제에서는 일어나는 두 대화자가 말을 주고 받는 횟수를 세었다. 실시간으로 상호작용하는 가운데 즉각적인 반응을 보여야 하는 대화 과제에서, 유창함에 대한 정보가 더 많이 드러날 수 있다는 것이 밝혀졌다. 설문지의 분석은 40명 중 38명의 시험자가 실제 대화 능력이 드러나기 때문에, 혼자 말하기 과제보다 대화 과제를 더 진정성있고, 실제적이며, 정확한 말하기 평가 수단이라고 인식했음을 보여주었다..

따라서, 본 연구는 상호작용적 전화대화 과제가 시험자의 말하기를 끌어내기 위해 혼자 말하는 과제로만 구성이 되어 있는 준직접적 말하기의 결함을 보완하기 위해, 상호작용 말하기 기술을 평가하는 신뢰할 수 있고 진정성 있는 측정방법임을 제시하였다. 여러 사회관계와 기능언어가 수반되는 전화 대화 상황 속, 다양한 목표언어 사용 영역에서 시험자의 수행능력과 인식을 연구하는 것이 필요하다.

주요어: 준직접 말하기 평가, 상호작용, 상호작용적 전화대화 과제, 롤플레이 과제, 목표 언어 사용 영역

학번: 2010-20020